NORTHWESTERN UNIVERSITY

Materials Discovery for Water Splitting Applications Using First-Principles Calculations and Machine Learning

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Materials Science and Engineering

by

Antoine A. Emery

EVANSTON, ILLINOIS

September 2017

ABSTRACT

Materials Discovery for Water Splitting Applications Using First-Principles Calculations and Machine Learning

Antoine A. Emery

Moving away from fossil fuels requires environmentally friendly and economically viable alternative energy sources. A wide adoption of new technologies for energy production and storage depends on better performing materials. Computational methods, such as electronic structure calculations and machine learning, hold the promise to work in conjunction with traditional experimentation to accelerate the discovery of materials needed to make those technologies more efficient. This thesis presents a first-principles methodology in the context of perovskite discovery for hydrogen fuel production via solar thermochemical water splitting. We calculate the properties of an exhaustive list of compounds to search for the ideal materials for water splitting, some that have never been experimentally synthesized. In addition, we use this large dataset of materials to benchmark current machine learning techniques to further reduce the number of expensive calculations required to discover new materials. Finally, we look at the entropy of reduction of cerium to explain the good performance of ceria for water splitting.

Acknowledgements

This work would not have been possible without the help and support of many people, both within and outside of the academic world.

First, Chris Wolverton, my advisor. Thanks for giving me the freedom and means to pursue my research in the direction I wanted. Thank you for your constant insightful scientific advice and pragmatism.

Thanks to the entire Wolverton research group, past and present, for their support, friendship and endless source of motivation. In particular, Vinay Hegde for being the "go-to" person for any questions, James Saal for teaching me the ropes of computational materials science, Maximilian Amsler for constant advice and Shahab Naghavi for showing me the chemistry side of every story.

To all my friends and colleagues at Northwestern University and in Chicago, thanks for making my graduate school and American experience unforgettable.

Last but not least, *un chaleureux merci* to my parents, brother and family for their continuous moral support, uninterrupted encouragement and for showing interest in my work. Thank you for allowing me to realize my own potential.

Abbreviations

BVM: bond valence method.

CF: crystal field.

DFT: density functional theory.

GB: gradient boosting.

GGA: generalized gradient approximation.

HT-DFT: high-throughput density functional theory.

ICSD: inorganic crystal structure database.

LDA: local density approximation.

MC: Monte Carlo.

ML: machine learning.

OCP: opposing crystal potential.

OQMD: Open Quantum Materials Database.

PAW: Projector-augmented wave.

PBE: Perdew-Burke-Ernzerhof.

RF: random forest.

SOC: spin-orbit coupling.

SVM: support vector machines.

TR: thermal reduction.

TWS: thermochemical water splitting.

TWSC: thermochemical water splitting cycles.

VASP: Vienna ab initio package.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

As a result of an ever-increasing number of power consuming devices, the world's energy demands are constantly rising. As this increase cannot be stopped and is necessary for economical development, there is a need for sustainable and renewable energy to power today's and tomorrow's society.[1,2] At the core of clean energy is a materials challenge: we need to discover new, or improved, materials that meet the requirement for environmentally friendly production of energy. Energy production technologies such as photovoltaic cells, thermoelectric materials, solid oxide fuel cells, and energy storage processes such as batteries and solar fuels are all active areas of research. While the performance of a new material can be assessed experimentally by synthesizing and directly testing materials, it is time consuming and expensive to perform on a large scale. On the other hand, electronic structure calculations, such as density functional theory, offer exceptional tools to perform such performance assessments. In addition to first-principles calculations, recent developments in materials informatics can be used to predict very cost-effectively certain properties of materials.

In this work, we use both density functional theory and machine learning to look at perovskites for hydrogen production via solar thermochemical water splitting. H_2 as a fuel presents several advantages: it has high energy density, can be used in mobile and stationary devices, and the product of its combustion is water. One carbon-free approach to produce hydrogen is to split the water molecule using thermal energy from the sun.

As water is a very stable molecule, doing so directly requires too high temperatures to be practically feasible. An alternative route is to use metal oxides to perform two-step thermochemical cycles. Despite success with ceria (CeO_2), the state-of-the-art material for this application, there is a need to discover new materials to increase the efficiency of the process for large scale hydrogen production. For that purpose, ABO_3 perovskites have shown some promise in reducing the operating temperatures. The remarkable stability of the perovskite structure with respect to its constituent elements and its ability to tolerate large amounts of oxygen vacancy without phase transformation, suggests that there could be potentially superior perovskites yet to be discovered.

In this thesis, we will look at efficient ways to screen perovskites for water splitting by using density functional theory and machine learning techniques. The structure of the thesis is shown in Figure 1.1. It highlights how a dataset of ABO_3 compounds is used for two different purposes 1) screening for water splitting materials and 2) machine learning optimization. The next chapter contains some background on water splitting, perovskites, density functional theory and machine learning. Chapter 3 will describe the high-throughput calculations framework and validate our results with experiments. In chapter 4, we use the calculations from chapter 3 to screen for thermodynamically favorable perovskites for thermochemical water splitting. Chapter 5 explains why ceria and cerium containing compounds are heavily featured for water splitting and redox applications in general. In chapter 6, we use the perovskite dataset of chapter 3 to benchmark several machine learning technique and develop models to reduce the need for expensive density functional calculations. Finally, chapter 7 summarizes and presents outlook for all the work presented in this thesis.

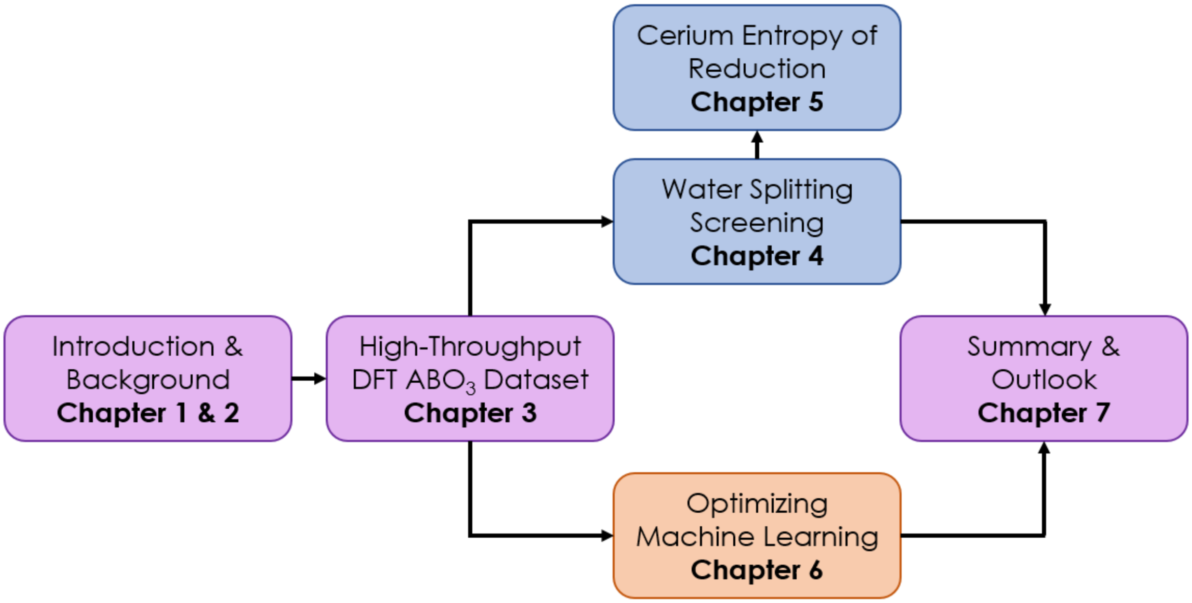

Cerium Entropy of Reduction
Chapter 5
Water Splitting Screening
Chapter 4
Introduction & Background
Chapter 1 & 2
High-Throughput DFT ABO$_3$ Dataset
Chapter 3
Summary & Outlook
Chapter 7
Optimizing Machine Learning
Chapter 6

Figure 1.1. Visual table of contents of the work presented in this thesis.

CHAPTER 2

Background

2.1. Water Splitting

The clean and sustainable production of energy is a major challenge faced by today's world. The increase of energy demand, the depletion of fossil fuels and the greenhouse gases emitted during their combustion call for new non-polluting and renewable energy sources. Hydrogen is a promising solution to address these issues as it has high energy density and the product of its combustion is water. Furthermore, it can be used for stationary and mobile applications by direct combustion or as fuel in solid oxide fuel cells.[3]

To produce hydrogen from water using solar power, three primary methods are possible: electrolytic, photochemical and thermochemical processes.[4] As these approaches involve different reaction mechanisms, a consistent way of measuring their efficiency is required. The solar-to-hydrogen production efficiency is often used for that purpose and is defined as:

$$\eta = \frac{m_{\mathrm{H}_2} * \Delta H_{\mathrm{H}_2}}{I_S} \tag{2.1}$$

where m_{H_2} is the amount of hydrogen produced in moles per second, ΔH_{H_2} is the upper heating value of hydrogen in joules per mole and I_s is the total solar irradiance of the solar spectrum, in units of joules per second.

With the electrolytic approach, the water molecule is dissociated by passing a direct current, coming from an external source, through two electrodes immersed in water. If the electric potential is high enough, hydrogen is produced at the cathode and oxygen at the anode. Solar power is used to create the current with photovoltaic (PV) panels or with a heat engine driven by solar-generated gases.[4] As the electrolysis process is technologically mature, the solar-to-hydrogen efficiency is mainly governed by the electricity generation from PV panels or heat engines. The reported solar-to-fuel efficiency for water electrolysis, using PV panels as source of current, is between 10% and 16%.[5] Using a solar thermal plant to generate steam or other gases in order to run heat engines can double this total efficiency.[4] In addition to the maturity of the electrolysis technique, another advantage of this route is the flexibility in current source: as long as it is produced in a clean way, current can be directly drawn form the grid.

With the photochemical route, metal oxides are irradiated with photons to excite electrons. Those excited electrons can then drive the reduction of water at the surface of the material.[6–8] Several metal oxides, including $AgNbO_3$, $PbTiO_3$, $CuNbO_3$ and $NaTaO_3$[8,9] have been successfully used to perform water splitting. The main advantages of this route are the absence of external power and the simplicity of the setup. However, as only a portion of light with an energy greater than the materials' band gaps is useful, the total efficiency is penalized. Typical values for the solar-to-fuel efficiency is in the order of 1%.[4,8]

2.1.1. Thermochemical Water Splitting

The thermochemical approach uses thermal energy to split water molecules. Thermochemical processes have the advantage over electrolytic and photochemical approaches of using the entire solar spectrum and of not requiring intermediate steps such as electricity production. As a result, thermochemical processes have potentially higher efficiency than other routes and is thus suitable for large hydrogen production plants.

Direct thermolysis of water, even if conceptually simple, requires too high temperatures and a gas separation mechanism, making it practically not feasible at a large scale.[10–12] Using direct thermolysis, at least 2500 K is required to achieve a reasonable degree of conversion and the Gibbs free energy equals zero under a pressure of 1 bar at a temperature of 4300 K.[13,14] In order to reduce the water splitting temperatures, thermochemical cycles catalyze the dissociation of water molecules using metal oxides to perform two, or more, redox reactions.

The idea of splitting water molecules using thermal energy was first explored in the 1960s, as a way to diversify the use of nuclear energy.[11,15] Therefore, it was characterized by the use of temperatures below 1223 K, which was available with nuclear reactors. At such temperatures, more than two chemical reaction steps are necessary, which proved to be challenging in terms of product separation and introduced inherent inefficiencies due to heat transfer.

The concept of using two-step water splitting cycles involving metal oxides came around 1977.[10] The general chemical reactions are presented in Figure 2.1. The top equation of Figure 2.1 is referred as the thermal reduction (TR) step that occurs at high temperature (typically around 2000 K) and the bottom equation of Figure 2.1 is the gas

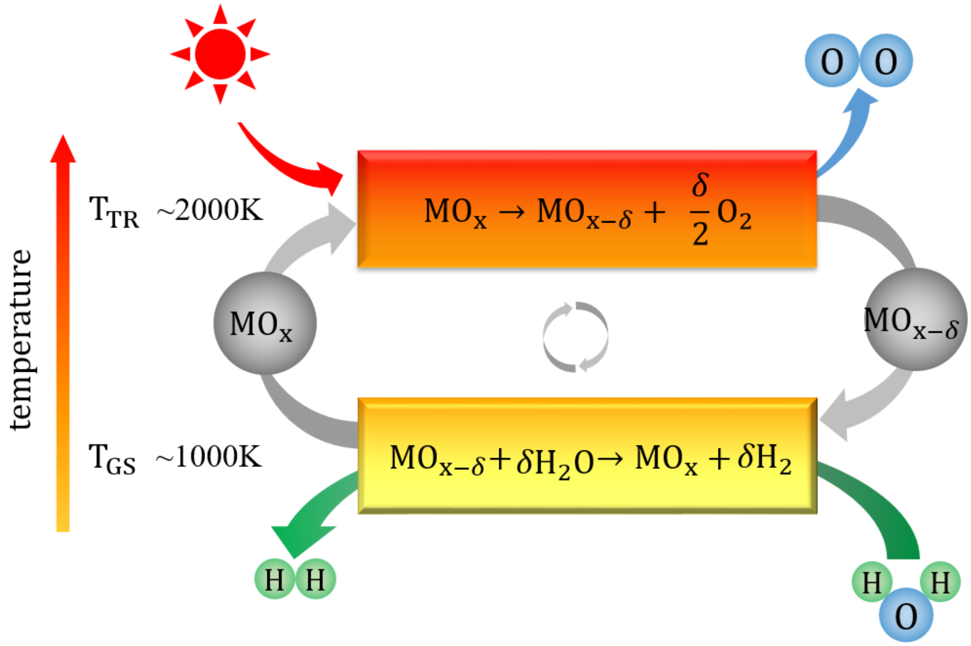

temperature
T_{TR} ~2000K
T_{GS} ~1000K
$MO_x \rightarrow MO_{x-\delta} + \frac{\delta}{2}O_2$
MO_x
$MO_{x-\delta}$
$MO_{x-\delta}+\delta H_2O\rightarrow MO_x+\delta H_2$
O O
H H
H H
O

Figure 2.1. Thermochemical water splitting process. The top equation is the thermal reduction, the bottom equation is the gas splitting reaction. M is a metal, MO_x its corresponding metal oxide and δ is the degree of oxygen off-stoichiometry.

splitting step occurring at a lower temperature (around 1000 K).[11] In Figure 2.1, δ is the oxygen off-stoichiometry, i.e. the amount of oxygen vacancy that is created when reducing the metal oxide. It becomes evident that maximizing δ is key to increase the efficiency of the water splitting process. The bigger the δ the higher the amount of H_2 is produced per mole of metal oxide. As a result, the overarching goal is to maximize the amount of oxygen vacancy, while avoiding phase transformation of the metal oxide as it would impact negatively the kinetics of the reactions.

The use of a metal oxide, that is recycled during the process, essentially solves the problems from the direct thermolysis (high temperature requirement) and the “more than two steps" thermochemical cycles (heat losses due to heat transfer). Indeed, two-step thermochemical cycles reduce the process temperature by about 800-1000 K while diminishing the energy losses during heat transfer. Furthermore, as oxygen and hydrogen are produced at different stages of the process, their recombination is avoided. Additionally, temperatures required for two-steps thermochemical water splitting can be reached using concentrated solar power, providing a means to produce hydrogen in an environmentally friendly way. As a consequence, efforts have been directed on metal oxides that would thermodynamically favor both TR and GS reactions. Around 30 thermochemical water-splitting cycles in a range of 1173-2273 K were discovered and tested.[3] The ones encountered frequently in the literature include zinc oxide (ZnO/Zn),[16,17] tin oxide (SnO_2/SnO),[18,19] iron oxide[10] and ceria[20–25] based cycles.

Iron oxide (Fe_3O_4/FeO) was the first redox cycle proposed by Nakamura *et al.*[10] However, the reduction proceeds above the melting point of Fe_3O_4 and FeO, resulting in loss of surface area due to the liquefaction of the material.[10,26] Different approaches to circumvent this problem were investigated. Among them, alloying metals into ferrite (MFe_2O_4 with M = Co, Ni and Zn)[27] and using yttrium-stabilized cubic zirconia to support the iron oxide[28] were proven to be, at least at a laboratory scale, feasible but kinetically slow.

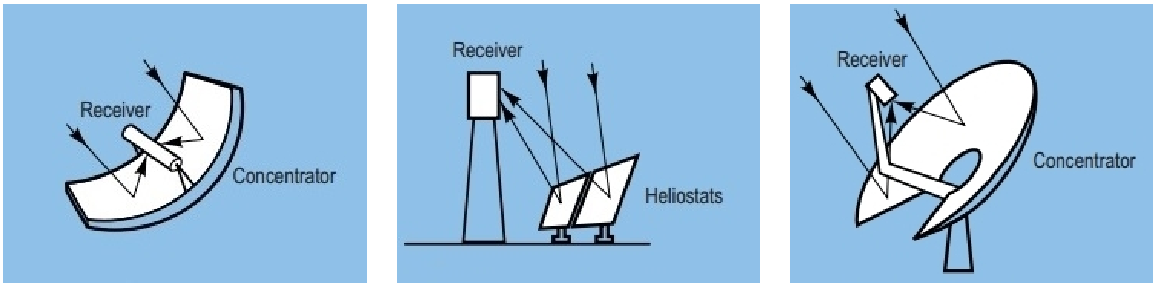
Ceria was proposed by Otsuka *et al.*[20] as a candidate for water splitting and was recently studied thoroughly as it appears to be one of the most promising material for this application as it can accommodate a large number of oxygen vacancy without phase transformation.[20–25,29,30] The absence of phase transformation impact positively the kinetics of the reaction and the stability of ceria. Indeed, fast kinetics, at 2273 K, were proven by

Abanades *et al.*[21] and ceria was cycled 500 times exhibiting stable oxygen release and hydrogen production after 100 cycles.[23] The main drawback of ceria based cycle is the need for high temperature during the thermal reduction step. This is impacts negatively the overall efficiency of the process as large temperature swings causes inevitable heat losses. To reduce heat losses, isothermal water splitting has been proposed.[12,31] In this setup, oxygen partial pressure is used to control its chemical potential instead of temperature.

Due to having a high tolerance to oxygen vacancy, lanthanum based perovskites have been experimentally used to perform solar thermochemical water splitting.[32–34] Doped $LaAlO_3$, showed promising results in term of oxygen release during thermal reduction (up to eight times larger than ceria).[33] Additionally, $La_{1-x}Sr_xMnO_3$ exhibited higher solar-to-fuel efficiency than ceria under low temperature condition.[34,35]

2.1.2. Concentrated Solar Power

There are three ways of concentrating solar power: trough, tower and dish systems (Figure 2.2).[36] The concentration factors of those setups are measured by $\tilde{C}$.



(a) trough (b) tower (c) dish

Figure 2.2. Three different ways of concentrating solar power. Concentration factors are equal to 100, 1000 and 10,000 for trough, tower and dish, respectively.[36]

$$\tilde{C} = \frac{Q_{\text{solar}}}{I * A} \tag{2.2}$$

In equation 2.2, Q_{solar} is the solar power received by the target, A is the targeted area and I is the incident normal beam insolation.[14] When I is normalized to 1000 $\mathrm{W\,m^{-2}}$, $\tilde{C}$ is expressed in "suns". It is on the order of 100, 1000, 10,000 for trough, tower and dish systems, respectively. When used as a chemical reactor, solar concentrating systems usually incorporate a cavity-receiver device. It consists of a insulated enclosure with a small opening to let in concentrated solar radiation. This configuration allows to approach the characteristics of a blackbody receiver. Following the derivation of Steinfeld *et al.*, the exergy efficiency, i.e. the efficiency of the conversion between solar and chemical energy stored in H_2, of an ideal thermochemical process is given by:[14]

$$\eta_{\text{exergy, ideal}} = \eta_{\text{absorption}} * \eta_{\text{Carnot}} = \left(1 - \frac{\sigma T^4}{I\tilde{C}}\right) * \left(1 - \frac{T_L}{T}\right) \tag{2.3}$$

where σ is the Stefan-Boltzmann constant ($5.67{*}10^{-8}\,\mathrm{W\,m^{-2}\,K^{-4}}$) and T_L is the temperature of the reservoir for heat rejection (usually room temperature). For instance, a typical tower setup, with $\tilde{C} = 1000$ and $T_{\text{optimal}} = 1105\,\mathrm{K}$, has a theoretical efficiency of 67%. In practice, as a result of convection, conduction loss and materials constraints, reported efficiencies are lower and range between 40% and 60%.[4,37]

2.2. Perovskites

In this work, we largely focus on perovskite materials. In chapter 3, we calculate an exhaustive dataset of such compounds that we will use in chapter 4 to screen for water splitting materials and in chapter 6 to optimize machine learning techniques. As such, we briefly describe the perovskite structure and its numerous applications in this section.

The ideal perovskite structure adopts a cubic symmetry and has a formula of ABX_3, where A and B are two different cations and X is an anion. As they are the most frequently encountered and are relevant for our water splitting application, we will focus on perovskite oxides, i.e. $X = O^{-2}$, for the entirety of this work. The cubic structure consists of a 12-fold coordinated A-site cation sitting in the center of a cube with B-site cations sitting on the corners of the cube. The B-site cations are octahedrally 6-fold coordinated by oxygen atoms. Those BO_6 octahedra are thus corner-sharing in all directions (see Figure 2.3 (a)).[38,39] Aside from the ideal cubic structure, where the B-atom and oxygen atoms are linearly arranged, the structure can undergo several distortions. The most commonly observed ones show tilting of the BO_6 octahedra (rhombohedral and orthorhombic distortion Figure 2.3 (b) and (d)) or displacement of the central A-atom (tetragonal distortion, 2.3 (c)).[40]

In an ideal cubic perovskite structure, the following relationship is true:

$$r_A + r_O = \sqrt{2}(r_B + r_O) \tag{2.4}$$

where r_A, r_B and r_O are the ionic radii of the A, B and oxygen atoms respectively. To quantify the deviation from the ideal cubic perovskite structure, Goldschmidt defined a tolerance factor as:[41]

$$t = \frac{r_A + r_O}{\sqrt{2}(r_B + r_O)} \tag{2.5}$$

The closer t is to 1, the closer to the ideal cubic structure the perovskite is.

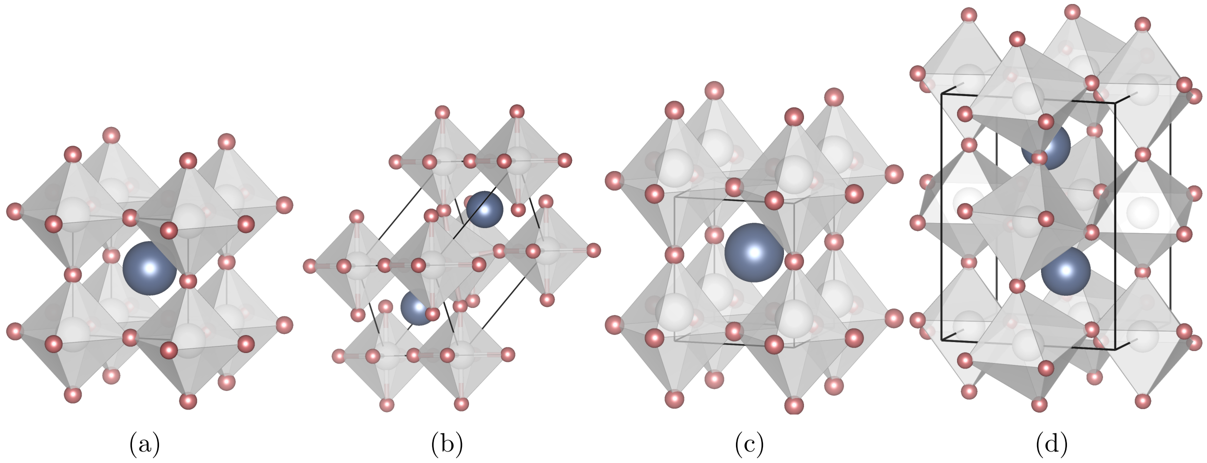


Figure 2.3. (a) Cubic perovskite structure (Pm$\bar{3}$m, 221). Blue atoms are the A-site cation, white atoms are the B-site cations that are embedded in oxygen (red atoms) octahedron. (b) Rhombohedral distortion (R$\bar{3}$c, 167). (c) Tetragonal distortion (P4mm, 99). (d) Orthorhombic distortion (Pnma, 62). The unit cell (black lines) of the cubic, rhombohedral, tetragonal and orthorhombic structure contains 5, 10, 5 and 20 atoms respectively. Distortions are exaggerated to be visible.

As a result of the numerous distortions and large number of elements that can be incorporated in the structure, perovskites have a large variety of magnetic, electronic and ionic properties which makes them invaluable for many technological applications such as dielectric, ferroelectric, piezoelectric, solid oxide fuel cells, photocatalysis, solar cells, and thermochemical water splitting.[38,39,42–47]

2.3. Quantum Mechanics and the Many-body Problem

Most of this work is based on our ability to compute energies of materials by solving efficiently electronic structures of many compounds. Physics at a small scale, such as electrons and atoms, is governed by quantum mechanics and the Schrödinger's equation.

The relevant form for materials science where electrons are in a steady state, is the time-independent Schrödinger's equation[48,49] (equation 2.6).

$$\hat{H}\Psi = E\Psi \tag{2.6}$$

where Ψ is a wave function, E is the energy of the state Ψ and $\hat{H}$ is the time-independent Hamiltonian. For systems containing multiple electrons, i.e. technologically relevant materials, the Hamiltonian is expressed by:

$$\hat{H} = -\frac{\hbar^2}{2m}\nabla^2 + V_{n-e}(\mathbf{r}) + V_{e-e}(\mathbf{r}) \tag{2.7}$$

where the first term on the right side of the equation is the kinetic energy, the second term is the nucleus-electron potential and the last term is the electron-electron interaction which makes equation 2.7 impossible to solve analytically. As a result, some approximations are required to use quantum mechanics for real systems.

2.3.1. Hartree-Fock

One way to solve the electron-electron interaction of equation 2.7 is to deal with $V_{e-e}(r)$ by considering independent electrons that are moving in an effective potential that represent the average repulsive interactions of the other electrons (mean-field theory).[50] In addition, the Hartree-Fock method[51,52] guarantees that the wave function is anti-symmetric by using a single Slater determinant.[53] Hartree-Fock is widely used in quantum chemistry but suffers from several problems: even though it includes the exact exchange due to the Pauli exclusion principle (by ensuring that the wave function is anti-symmetric), it neglects any correlations between electrons. In addition, Ψ is still a function of 3N variables (where N is the number of electrons).

2.3.2. Density Functional Theory (DFT)

Another way of dealing with the many-body problem came when Hohenberg and Kohn showed that the ground-state energy of the Schrödinger equation is a unique functional of the electron density, $n(\mathbf{r})$,[54] thus reducing the dimensionality of the problem from 3N to 3. Hohenberg and Kohn also showed that the external potential of a system is uniquely determined by the ground-state charge density and that the total energy of a system can be written as an unknown functional of the ground-state charge density. This functional ($F[n(\mathbf{r})]$) is subject to the variational principle meaning that the ground-state charge density is the one that minimizes the total energy of the system.

As the energy of an interacting electron gas is still unknown and unsolvable analytically, Kohn and Sham proposed to transform the problem into a simpler problem.[55] In a Kohn-Sham system, the interacting electrons are replaced by non-interacting particles evolving in an external potential. Doing so allows the separation of the known terms (kinetic energy, T_s, and Hartree energy, E_H) from the unknown part, the exchange-correlation energy (E_{xc}). The functional is thus defined as:

$$F[n(\mathbf{r})] = T_s[n(\mathbf{r})] + E_H[n(\mathbf{r})] + E_{xc}[n(\mathbf{r})] \tag{2.8}$$

The exchange-correlation functional contains two quantum effects that are not included in single-particle Kohn-Sham systems: the difference in energy due to the Pauli exclusion principle (exchange) and the interaction between electrons of opposite spins (correlation).

The accuracy of density functional theory (DFT) depends mainly on the approximation of the exchange-correlation function. Two ways of approximating it are widely used: the local density approximation (LDA) and the generalized gradient approximation (GGA).[56,57]

In practice, DFT is an iterative method where the ground-state charge density is solved self-consistently. 1) An initial guess of the charge density is made 2) an effective potential is constructed 3) a new charge density is calculated from the resulting wave functions. The process is repeated from step 2) until the initial and final charge densities are the same. The outputs of a DFT calculation are the total energy of the system, the forces on the atoms (which can be used to relax the structure), the ground-state charge density and the independent-particle wave functions and eigenvalues.

2.3.3. High-Throughput Density Functional Theory (HT-DFT)

The versatility of density functional theory coupled with the increasing performance of computers enable high-throughput calculations, where large numbers of calculations (typically thousands) are performed.[58] As a result of this increase of computational power, several databases of calculations, encompassing ten thousands of compounds were created.[59–62] In order to compare total energies of all the compounds, such databases require consistent settings for all the calculations. This often comes at a price of reduced accuracy as certain properties, such as complex magnetic structures, have to be neglected. However, those databases offers a way to screen for certain properties in large dataset of compounds. This methodology has been successfully tested in the past for a variety of applications.[63–68]

The majority of the calculations from this work are done using the Open Quantum Materials Database (OQMD) a framework that was developed in the Wolverton group.[62] It includes consistent calculations of structures present in the inorganic crystal structure database (ICSD)[69,70] as well as several prototypes, i.e. theoretical structures that were not necessary experimentally observed. Each compound is relaxed to find the optimal configuration of atoms with the lowest energy.

2.4. Machine Learning

Machine learning algorithms can learn and make predictions based on data without being explicitly programmed to do so. It is employed in a variety of fields including computer vision,[71] fraud detection,[72] speech recognition,[73] biology,[74] and materials science.[75] Machine learning algorithms are classified in two categories: supervised and unsupervised algorithms. Unsupervised algorithms are used to identify pattern in data. Clustering data points in different categories is on example of unsupervised learning.[76]

Most algorithms used in materials science are supervised algorithms where models are build to map the inputs (feature sets or attributes) of an unknown process to its outputs (target values). Typically, a model is trained on dataset of known points, i.e. where both the feature sets and target values are known (training set) and then used to predict the properties of another set of data where only the feature sets is known (unseen set). Supervised learning can be as simple as linear regression or polynomial fitting but also includes more complex algorithms such as neural network and ensemble methods.[77–79]

Different algorithms have different advantages and disadvantages which can depend on the type of dataset that are being studied. A key concept to choose the right algorithm is the “generalization error” or “out-of-sample error” which measures the ability of the model to predict the outcome of the unseen set. The generalization error is comprised of the irreducible error, corresponding to the noise in the dataset, the bias, and the variance.[78] The last two terms are commonly referred as the “bias-variance trade-off” (or dilemma). It arises because minimizing the error on the training set (bias) increases the error due to small fluctuations in the data (variance). Indeed, underfitting (high bias) can cause the algorithm to miss relevant relation between the inputs and the outputs, whereas overfitting (high variance) increases the error of the prediction on the unseen set. Qualitatively, finding a complex model to represent accurately the training set can lead to high variance in the prediction (overfitting) whereas finding a model that is too simple might not capture all the characteristics of the data (underfitting). The performance of the model is often measured in function of the n-fold cross-validation score (where n is typically 10), i.e. partitioning the training set in n part, training a model on n-1 partition, using the model to predict the data points that were excluded from the training set and repeating this process n times.

Apart from the different algorithms, the generation of inputs, i.e. feature sets or attributes, plays a critical role in the adoption of machine learning for materials science. In short, we have to choose how to represent materials to the algorithm, generally expressed as a list of numerical values. These representations have to fulfill several requirements such as being able to distinguish materials from each other while capturing the relevant physics of the compounds but also simple enough so that computing the representations is faster than the methods used to generate the training set. Different approaches have been used in the literature to represent materials, some are composition based with a small number of attributes[80,81] while others aim at designing a framework broadly applicable to many materials and properties.[82] Generating those representations is an active area of research.[83,84]

CHAPTER 3

High-Throughput DFT Calculations of Perovskites

ABO_3 perovskites are oxide materials that are used for a variety of applications such as solid oxide fuel cells, piezo-, ferro-electricity and water splitting. Owing to their remarkable stability with respect to cation substitution, new compounds for such applications potentially await discovery. In this work, we present an exhaustive dataset of formation energies of 5,329 cubic and distorted perovskites that were calculated using first-principles density functional theory. In addition to formation energy, several additional properties such as oxidation states, band gap, oxygen vacancy formation energies, and thermodynamic stability are also made publicly available. This large dataset for this ubiquitous crystal structure type contains 395 perovskites that are predicted to be thermodynamically stable, of which many have not yet been experimentally reported, and therefore represent theoretical predictions. The dataset thus opens avenues for future use, including materials discovery in many research-active areas.

3.1. Background

As a result of their large tolerance to oxygen vacancy, ABO_3 perovskites are widely used for a variety of applications such as solid oxide fuel cells, piezo-, ferro-electricity and thermochemical water splitting.[38,39,43] Furthermore, their remarkable structural stability with respect to their constituent elements suggests that potential new compounds remain to be discovered. As the number of possible ABO_3 compounds is large, we use high-throughput density functional theory (HT-DFT) to compute the thermodynamical stability of 5,329 compositions in an exhaustive manner. In addition to the compounds stability, we also calculate the oxygen vacancy formation energies as it is a relevant quantity for many applications involving reduction of these compounds.[38,39,43]

All the 5,329 compounds are created by substituting 73 metals and semi-metals of the periodic table of the elements (see Figure 3.1) on the A and B sites (73^2 = 5,329) of the ABO_3 perovskite crystal structure. The ideal ABO_3 cubic perovskite crystal structure is composed of a B cation that is octahedrally 6-fold coordinated with oxygen atoms and an A cation that is 12-fold coordinated by oxygen atoms. Aside from the ideal cubic structure, many perovskites undergo a local distortion from this cubic structure; these distorted perovskites can have a variety of symmetries, including rhombohedral, tetragonal and orthorhombic distortions[39] (see Figure 2.3). In this work, all 5,329 compositions are calculated in the ideal cubic structure and a subset of those are calculated in the three aforementioned distortions.

The $T = 0\,K$, $P = 0\,bar$ ground state stability of all ABO_3 compounds was assessed with respect to all possible linear combinations of phases present in the A-B-O ternary phase diagram using a convex hull construction. All the phases that are used for the stability calculation are from the OQMD[62] and (as of July 2017) include ≈ 40,000 phases from the ICSD[69,70] and ≈ 430,000 hypothetical compounds based on decoration of common structural prototypes. The oxygen vacancy formation energy was calculated by using an $A_2B_2O_5$ structure, which corresponds to two perovskite unit cells with an oxygen atom removed. Additionally, other properties readily available from DFT calculations are reported, including the relaxed structure, band gap, and total magnetic moment. Figure 3.2 shows the workflow used to obtain all the quantities.

The present dataset will be used in a study aiming at identifying suitable perovskites for thermochemical water-splitting applications using both the stability and oxygen vacancy formation energy as screening parameters, see next chapter and ref. 68. This data is a valuable more generally in guiding experimental synthesis of predicted new compounds, further screening for a large variety of applications (other than water splitting) or to train machine learning (ML) models. While machine learning on materials dataset is an area of active research,[85–87] the datasets used by various research groups are often vastly different from one another with no way to compare various ML models. Having a large, consistent materials dataset that can be used by a variety of research groups to train machine learning models will allow a more transparent comparison of various methods being used in the field.

3.2. Methodology

3.2.1. Density Functional Theory

All DFT calculations were performed using the Vienna ab initio package (VASP).[88,89] Projector-augmented wave (PAW) potentials[90] are used with the Perdew-Burke-Ernzerhof (PBE)[91] generalized gradient approximation to the exchange-correlation functional. To improve the description of localized charge density of some 3d transition metals and most actinides, DFT+U is used (see Table 3.1, U is applied on d-electrons for transition metals and f-electrons for actinides).[92,93] Any calculations containing actinides or 3d elements (Sc-Cu) are spin-polarized with ferromagnetic alignment of spins. We note that this approach does not capture antiferromagnetism which is present in certain perovskites.[94] Based on a study by Stevanovic *et al.*, who calculated ternary compounds with up to ten

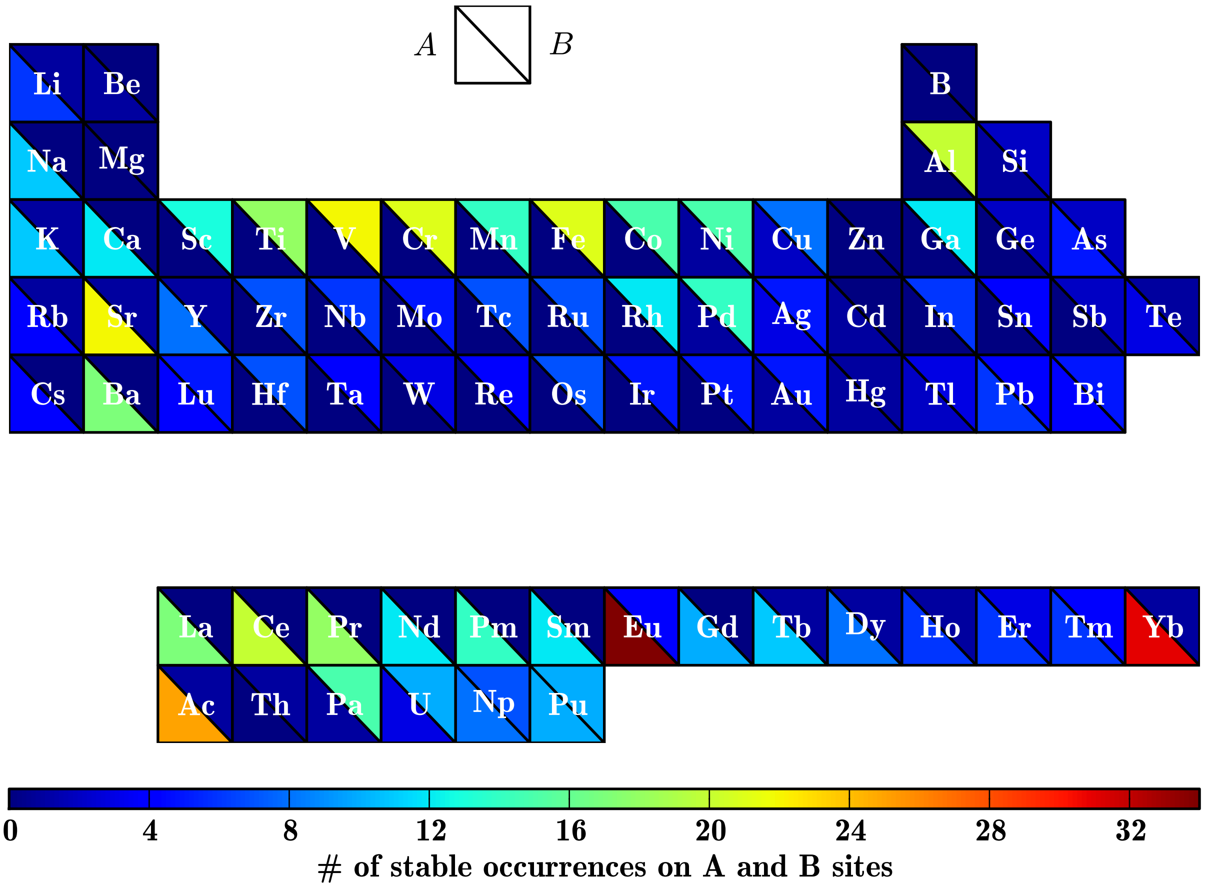


Figure 3.1. List of elements considered for the A and B sites. Elements are color-coded as a function of the number of stable perovskites predicted by DFT with the respective elements on the A and B sites.

different relative spin orientations, this computational error was found to be of the order of 0.01-0.02 eV/atom and thus is considered negligible for the purpose of this paper.[95] Calculations are performed through the OQMD framework.[62,96] The OQMD contains the energies of over 470,000 compounds comprising ∼40,000 structures from the ICSD[69,70] as well as more than 430,000 theoretical prototype structures. Calculation settings are explained in more detail in Kirklin *et al.*[62]

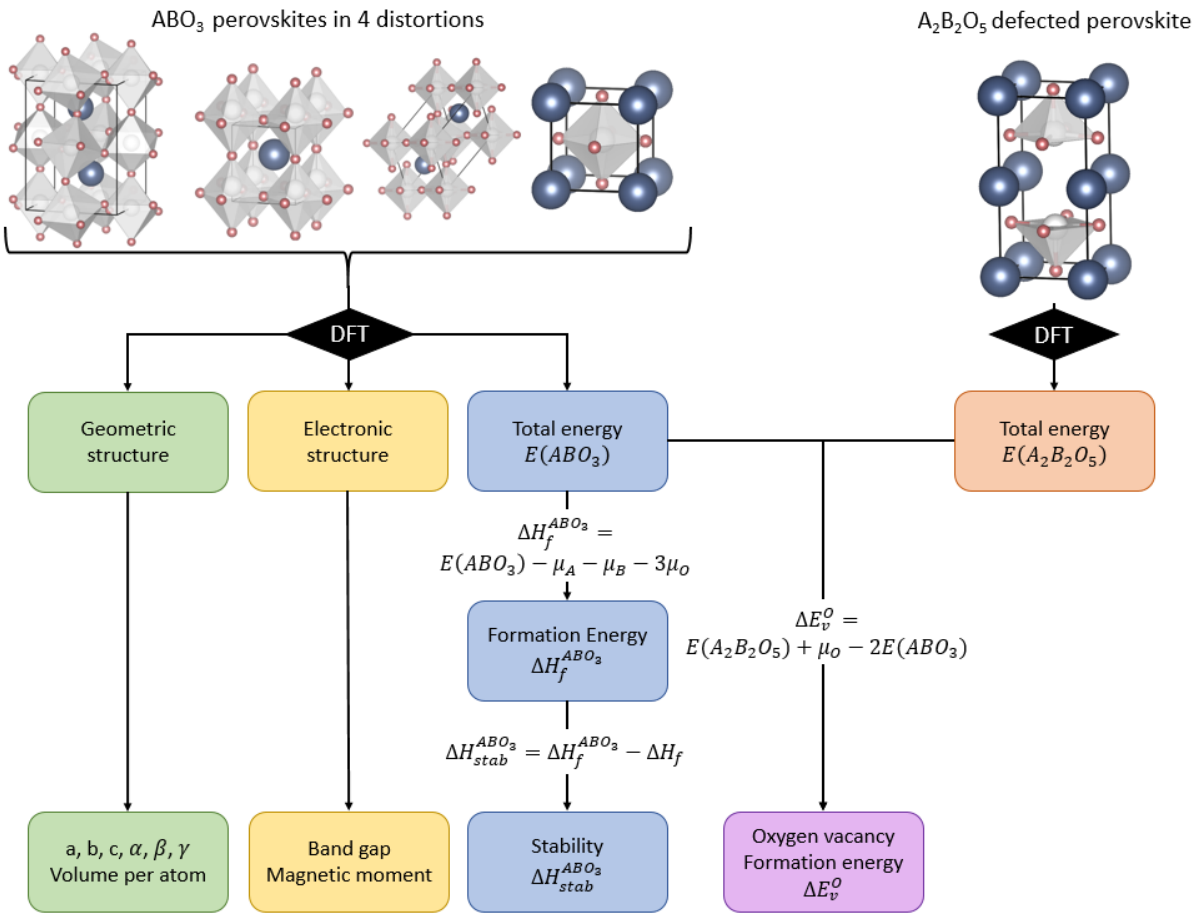

ABO_3 perovskites in 4 distortions
$A_2B_2O_5$ defected perovskite
DFT
DFT
Geometric structure
Electronic structure
Total energy $E(ABO_3)$
Total energy $E(A_2B_2O_5)$
$\Delta H_f^{ABO_3} = E(ABO_3) - \mu_A - \mu_B - 3\mu_O$
Formation Energy $\Delta H_f^{ABO_3}$
$\Delta E_v^O = E(A_2B_2O_5) + \mu_O - 2E(ABO_3)$
$\Delta H_{stab}^{ABO_3} = \Delta H_f^{ABO_3} - \Delta H_f$
a, b, c, α, β, γ Volume per atom
Band gap Magnetic moment
Stability $\Delta H_{stab}^{ABO_3}$
Oxygen vacancy Formation energy ΔE_v^O

Figure 3.2. Workflow to calculate all the properties in the current dataset. (top left) We start with all the cubic structures and compute all their total energies using density functional theory (DFT). If the stability (equation 3.1) of the cubic perovskite is less than 0.5 eV/atom (i.e., the cubic phase is within 0.5 eV/atom of the ground state convex hull), we also compute 3 additional distortions (orthorhombic, tetragonal, rhombohedral). The geometric properties (lattice parameters, angles, and volume per atom) and electronic properties (band gap and magnetic moment) are readily available from the calculations. Formation energies are calculated using elemental chemical potentials and thermodynamic stability is calculated with respect to all the other A-B-O phases present in the OQMD. (top right) Defected perovskites, 2x1x1 supercells with a missing oxygen atom, are calculated using DFT and their total energies, in conjunction with those of pristine cubic cells, are used to compute the oxygen vacancy formation energies of every composition.

Table 3.1. Chemical potentials and U-values used for the high-throughput density functional theory calculations.

Element	chemical potential [eV/atom]	U-value [eV]
Li	-1.897	-
Be	-3.755	-
B	-6.678	-
O	-4.523	-
Na	-1.199	-
Mg	-1.543	-
Al	-3.746	-
Si	-5.425	-
K	-1.097	-
Ca	-1.978	-
Sc	-6.328	-
Ti	-7.698	-
V	-6.263	3.1
Cr	-6.712	3.5
Mn	-6.940	3.8
Fe	-6.063	4.0
Co	-5.016	3.3
Ni	-2.999	6.4
Cu	-2.258	4.0
Zn	-1.266	-
Ga	-3.032	-
Ge	-4.624	-
As	-4.652	-
Rb	-0.963	-
Sr	-1.683	-
Y	-6.464	-
Zr	-8.547	-
Nb	-10.094	-
Mo	-10.848	-
Tc	-10.361	-
Ru	-9.202	-
Rh	-7.269	-
Pd	-5.177	-
Ag	-2.822	-
Cd	-0.900	-
In	-2.720	-

Table 3.1 – continued

Element	Chemical potential [eV/atom]	U-value [eV]
Sn	-3.914	-
Sb	-4.118	-
Te	-3.142	-
Cs	-0.855	-
Ba	-1.924	-
Lu	-4.524	-
Hf	-9.955	-
Ta	-11.853	-
W	-12.960	-
Re	-12.423	-
Os	-11.226	-
Ir	-8.855	-
Pt	-6.056	-
Au	-3.267	-
Hg	-0.359	-
Tl	-2.359	-
Pb	-3.704	-
Bi	-4.039	-
La	-4.935	-
Ce	-4.777	-
Pr	-4.775	-
Nd	-4.763	-
Pm	-4.745	-
Sm	-4.715	-
Eu	-1.888	-
Gd	-4.655	-
Yb	-1.513	-
Dy	-4.602	-
Ho	-4.577	-
Er	-4.563	-
Tm	-4.475	-
Yb	-1.513	-
Ac	-4.106	-
Th	-6.346	4.0
Pa	-9.496	4.0
U	-8.717	4.0
Np	-10.162	4.0
Pu	-12.087	4.0

3.2.2. Assessment of T = 0 K Stability

To determine, for a given A-B-O system, which phases are stable it is necessary to find, as a function of composition, the set of phases that have an energy lower than any other structures or any other linear combinations of structures. These ground-state phases are then linked with tie lines to form a *convex hull* (see Figure 3.3). The stability of a perovskite is given in terms of the energy difference between the perovskite and the convex hull, also referred as the hull distance, is defined as:

$$\Delta H_{\text{stab}}^{\text{ABO}_3} = \Delta H_f^{\text{ABO}_3} - \Delta H_f \tag{3.1}$$

where ΔH_f is the convex hull energy at the ABO_3 composition (not including the perovskite under consideration). Stable perovskites that are already reported and already in the OQMD will have a stability energy of zero (equation 3.1). Stable perovskites that are discovered during the course of this work will have a negative stability energy (equation 3.1) and perovskites with a positive value of $\Delta H_{\text{stab}}^{\text{ABO}_3}$ are unstable. Finally, $\Delta H_f^{\text{ABO}_3}$ is the DFT formation enthalpy of the perovskite calculated as follows:

$$\Delta H_f^{\text{ABO}_3} = E(\text{ABO}_3) - \mu_A - \mu_B - 3 * \mu_O \tag{3.2}$$

where $E(\text{ABO}_3)$ is the DFT total energy of the ABO_3 compound, μ_A, μ_B and μ_O are the chemical potentials of the A, B and oxygen species respectively. In most cases, μ_A and μ_B are DFT 0 K total energies of the crystalline elements. However, for elements exhibiting a phase transition between 0 K and 300 K (Na, Ti, Sn, O, Hg) and elements with DFT+U correction (V, Cr, Mn, Fe, Co, Ni, Cu, Th, Pa, U, Np, Pu), chemical potentials are fitted

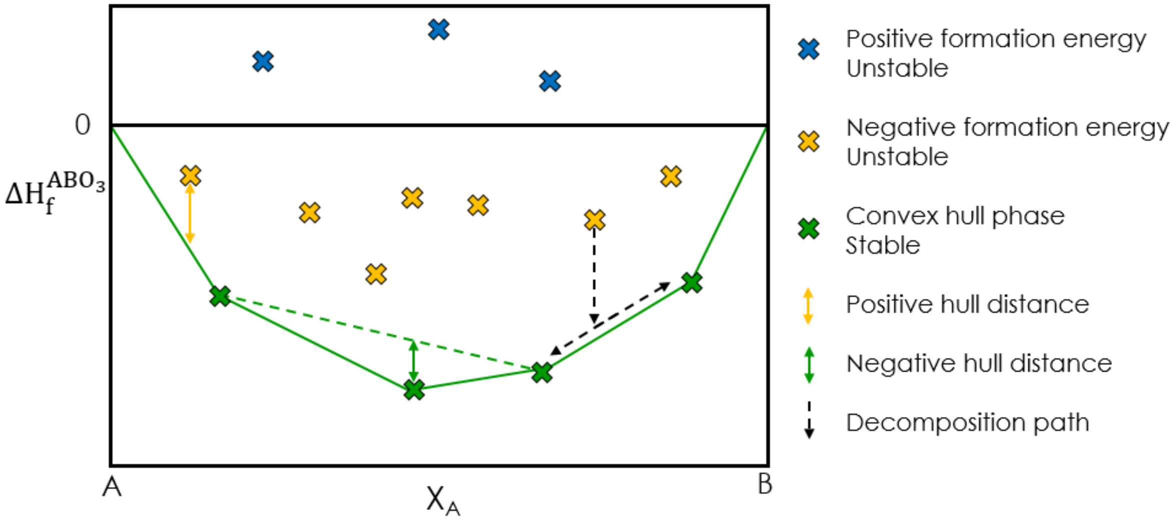


Figure 3.3. Schematic of a convex hull. $\Delta H_f^{\mathrm{ABO_3}}$ is the formation energy calculated by equation 3.2. The convex hull is formed by the green tie-lines and phases. Even though the yellow phases have a negative formation energy, they are not stable and will decompose in a linear combination of convex hull phases (black arrows).

to experimental values as detailed in ref. 62 (see Table 3.1). In particular, the oxygen chemical potential (μ_{O}) is fitted to experimental oxide ΔH_f values. Extensions of the grand canonical linear programming method (GCLP)[97] are used to construct convex hulls for every calculated system.[62,97] All 470,000 phases present in the OQMD are included in the stability calculation.

3.2.3. Structural Distortions of Perovskite

Aside from the ideal cubic structure (Figure 2.3 (a)), perovskites can exhibit several distortions.[39] The most frequently encountered are rhombohedral, tetragonal and orthorhombic distortions[39] (Figure 2.3 (b), (c) and (d)). Without experimental data for a given composition, or without using costly crystal structure prediction tools, it is not possible, a priori, to know what is the most stable structure of any given compound. As calculating every distortion for all the 5,329 stoichiometries would be computationally costly, we started by calculating the 5,329 different compounds in the undistorted cubic perovskite structure. Subsequently, we investigated the effect of distortions on the stability of perovskites by randomly selecting one-third of the compositions (1,776) and by computing their stability in the rhombohedral, tetragonal and orthorhombic distortion. We saw that distortions generally lower the energy of the ideal cubic structure but found no case where the distorted compound was lower in energy than the cubic phase by more than 0.5 eV/atom. Thus, we only calculated the distortion of compositions having a cubic stability lower than 0.5 eV/atom. This resulted in 2,162 $(1,776 + 386)$ compositions where the four distortions (cubic rhombohedral, tetragonal and orthorhombic) where calculated.

3.2.4. Oxygen Vacancy Formation Energy

For a general case, the oxygen vacancy formation energy, per vacancy, is calculated as follows:

$$\Delta E_v^{\mathrm{O}} = \frac{1}{\delta} E(\mathrm{ABO}_{3-\delta}) + \mu_{\mathrm{O}} - \frac{1}{\delta} E(\mathrm{ABO}_3) \tag{3.3}$$

where δ is the oxygen off-stoichiometry, $E(\mathrm{ABO}_3)$ is the DFT total energy of the defect-free cubic perovskite cell and μ_O is the same chemical potential as used in equation 3.2. Finally, $E(\mathrm{ABO}_{3-\delta})$ is the DFT total energy of a cubic supercell containing an oxygen vacancy. In the dilute limit, ΔE_v^O in equation 3.3 should be converged as a function of supercell size, until there is negligible interaction between the vacancy and its periodic images. However, large supercells are too computationally costly for our high-throughput approach. Hence, we use small supercell $A_2B_2O_5$ structures, which correspond to a 2x1x1 supercell size, and have a small unit cell (9 atoms) which enables the calculation of the oxygen vacancy formation energy for all perovskites in a high-throughput manner. Additionally, Curnan & Kitchin, who calculated oxygen vacancy formation energies for $LaBO_3$ and $SrBO_3$ (B = Sc-Cu) with different degrees of simplification, showed that trends in the oxygen vacancy formation energy are largely unaffected by the supercell size.[98] We confirmed this finding by performing calculations of several $LaBO_3$ perovskite vacancy formation energies for various supercell sizes, as shown in Figure 3.4.

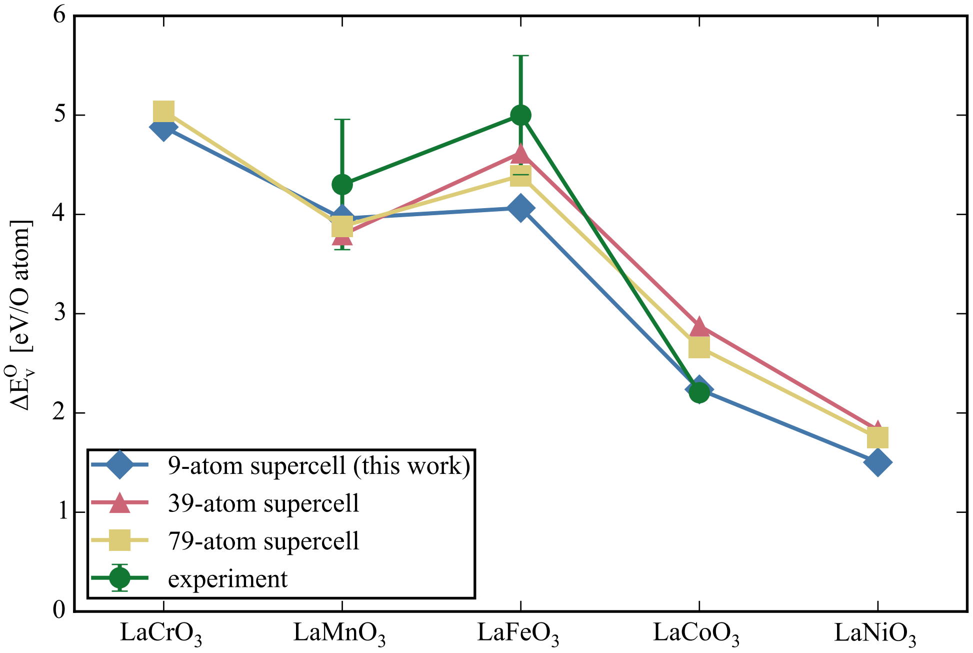


Figure 3.4. Comparison of oxygen vacancy formation energy calculated, in a high-throughput way, in the present work with oxygen vacancy formation energies calculated with larger supercells. Data are taken from Lee *et al.*[94] for the 40-atoms supercells and Deml *et al.*[99] for the 80-atom supercells. Experimental data are taken from Kuo *et al.*[100] and Nowotny *et al.*[101] for $LaMnO_3$, Mizusaki *et al.*[102] for $LaFeO_3$ and Mizusaki *et al.*[103] for $LaCoO_3$.

3.2.5. Oxidation States and Ionic Size

To keep our high-throughput methodology exhaustive, we did not restrict any compounds based on their charge balance or oxidation state prior to performing the calculations. However, to compare our stability results with previous study as well as drawing structure maps, it is necessary to estimate the oxidation state of A and B atoms in the perovskite compositions. For that purpose, we used a bond valence method (BVM)[104] to obtain the oxidation states of the cations for each of the 5,329 compositions that we studied. By fixing the oxidation state of oxygen to -2, we calculated the oxidation state of both cations using a BVM as implemented in pymatgen.[105] Nine elements used in our study (Tc, Os, Pt, Au, Pm, Ac, Pa, Np and Pu) do not have bond valences. Thus, compounds containing these elements are labeled as having an unknown charge state. With the oxidation states and the coordination numbers (12, 6 and 2 for the A-, B-atom and oxygen, respectively) of all the atoms in the structure, we used Shannon radii,[106,107] which were tabulated by Seshadri and Basu,[108] to estimate the size of each elements.

3.2.6. Data Records

The list of 5,329 ABO_3 perovskites can be found on figshare.[109] All the calculations, along with all the 470,000 compounds used for the stability calculations are available for download or for direct consultation at www.oqmd.org. Ref. 62 also contains detailed information about the calculation parameters. The data is stored in a CSV spreadsheet. Each row contains a different composition and each column is a property of that composition (described in Table 3.2). A calculation that did not converge to a final solution is indicated by a hyphen ("-") in the table for that composition. Those cases can happen for several reasons like having a bad initial structure requiring too many steps to converge or containing an element with a pseudo-potential that is hard to converge (notably the cesium pseudo-potential). This kind of computational issues is inherent to high-throughput methods where *consistent settings* have to be used for the calculation of *all* compounds in a reasonable amount of time.

Table 3.2. Description of column keys in the CSV spreadsheet hosted on Figshare.[109]

Name	Type	Unit	Description
Chemical formula	string	None	Chemical composition of the compound. The first and second elements correspond to the A- and B-site, respectively. The third element is always oxygen
A	string	None	Chemical element on the A-site
B	string	None	Chemical element on the B-site
Experimentally reported	boolean	None	Report of experimental synthesis of compound in the literature. True indicates that the compound is present in one of the four review papers.

Table 3.2 – continued

Name	Type	Unit	Description
Valence A	number or string	None	Valence of atom A as estimated by bond valence (BV) theory. If a compound is not balanced, it is denoted by not balanced. If the compound contains a least one element without a BV parameter, it is denoted by element not in BV
Valence B	number or string	None	Valence of atom B as estimated by bond valence (BV) theory. If a compound is not balanced, it is denoted by not balanced. If the compound contains a least one element without a BV parameter, it is denoted by element not in BV
Radius A	number	Å	Shannon ionic radius of atom A. When possible, the oxidation state and coordination number (12) of the A atom was used to estimate its radius.
Radius B	number	Å	Shannon ionic radius of atom B. When possible, the oxidation state and coordination number (6) of the A atom was used to estimate its radius.

Table 3.2 – continued

Name	Type	Unit	Description
Lowest distortion	string	None	Distortion with the lowest energy (among cubic, rhombohedral, tetragonal and orthorhombic corresponding to space group 221, 167, 99 and 62, respectively)
Formation energy	number	eV/atom	Formation energy as calculated by equation 3.2 of the distortion with the lowest energy
Stability	number	eV/atom	Stability (hull distance) as calculated by equation 3.1 of the distortion with the lowest energy. A compound is considered stable if it is within 0.025 eV/atom of the convex hull
Magnetic Moment	number	μ_B	Resulting magnetic moment of the relaxed structure. If the composition does not contain any magnetic element, the magnetic moment is set to a hyphen ("-").
Volume per atom	number	Å^3/atom	Volume per atom of the relaxed structure
Band gap	number	eV	PBE band gap obtained from the relaxed structure
a	number	Å	Lattice parameter a of the relaxed structure
b	number	Å	Lattice parameter b of the relaxed structure

Table 3.2 – continued

Name	Type	Unit	Description
c	number	Å	Lattice parameter c of the relaxed structure
alpha	number	°	α angle of the relaxed structure. $\alpha = 90$ for the cubic, tetragonal and orthorhombic distortion. α angle of the relaxed structure. $\alpha = 90$ for the cubic, tetragonal and orthorhombic distortion.
beta	number	°	β angle of the relaxed structure. $\beta = 90$ for the cubic, tetragonal and orthorhombic distortion.
gamma	number	°	γ angle of the relaxed structure. $\gamma = 90$ for the cubic, tetragonal and orthorhombic distortion.
Vacancy energy	number	eV/O atom	Oxygen vacancy formation energy as calculated by equation 3.3

3.2.7. Graphical Representation of the Data

The top part of Figure 3.1 shows the number of stable perovskites as a function of the elements occurring on the A- and B-sites. Out of 73 elements, only boron does not appear in any stable perovskites. Lanthanides and alkaline earths are frequently on the A-site for stable perovskites whereas transition metals, specially the first row, are comon on the B-site. Both those observations generally agrees with perovskites that are experimentally reported.[38,110] Figure 3.5 shows the formation energy and band gap distribution for all the compounds calculated in this work.

3.3. Technical Validation

In this section, we give different comparison with experiment and literature of our high-throughput calculations.

3.3.1. The Open Quantum Materials Database

The Open Quantum Materials Database uses DFT to compute the total energies of every compound. DFT is widely used in solid states physics due to its accuracy and reproducibility.[111–113] In addition, previous studies have shown that formation energies calculated using DFT, when compared against those measured experimentally, have a similar accuracy as a comparison between experimental values from two different sources.[62]

3.3.2. Lattice Parameters

For all the compounds that are predicted to be stable and have an entry in the ICSD, we compared the lattice parameters of the DFT relaxed structure with the lattice parameters of the experimental structure (Figure 3.6). The mean error (ME), mean absolute error (MAE), mean relative error (MRE) and mean absolute relative error (MARE) across all lattice parameter for the 113 compounds are 0.011Å, 0.048Å, 0.19% and 0.82%, respectively. The magnitude and overestimation of the lattice parameters are consistent with other lattice parameters studies in the literature for DFT-PBE.[114]

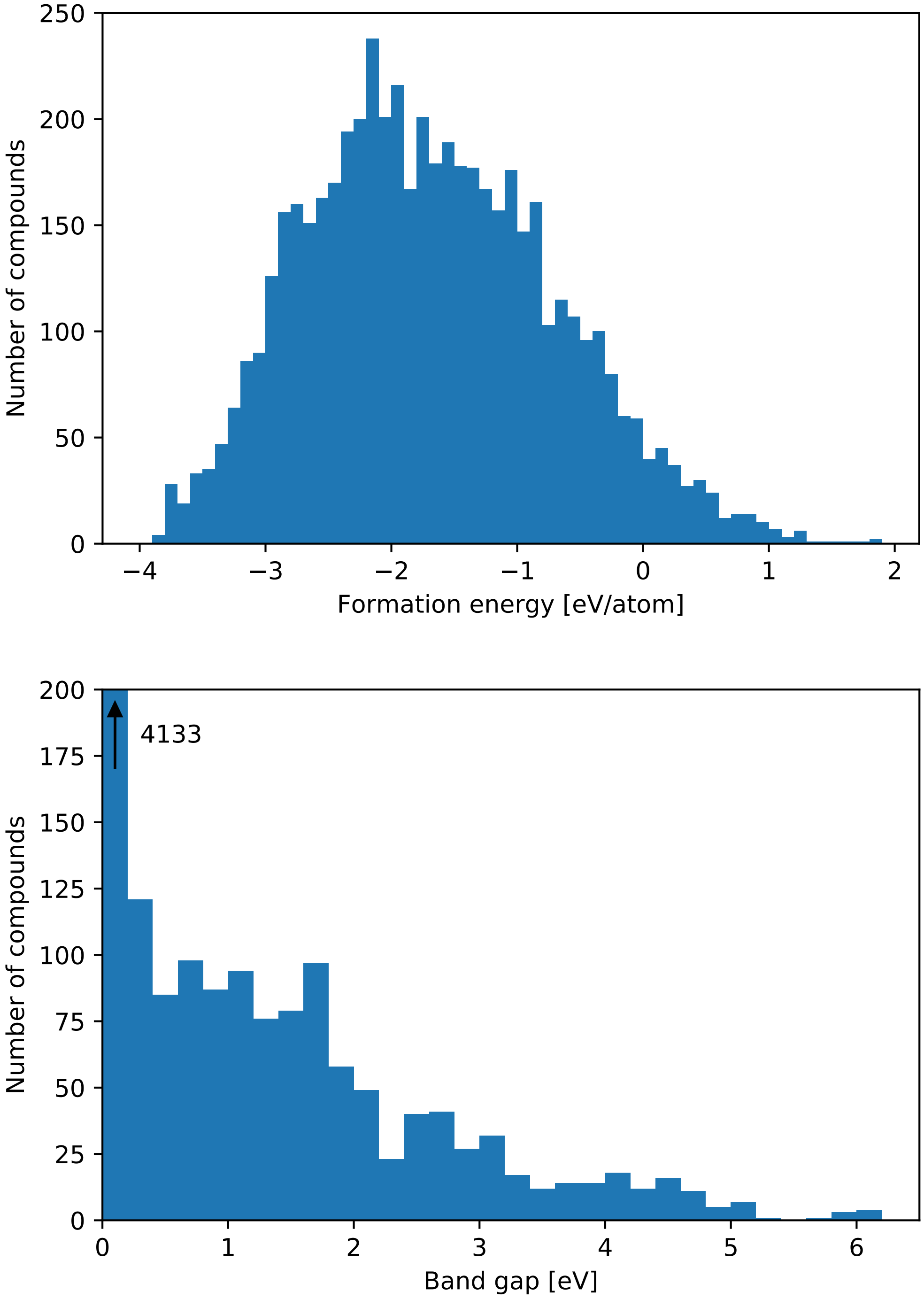
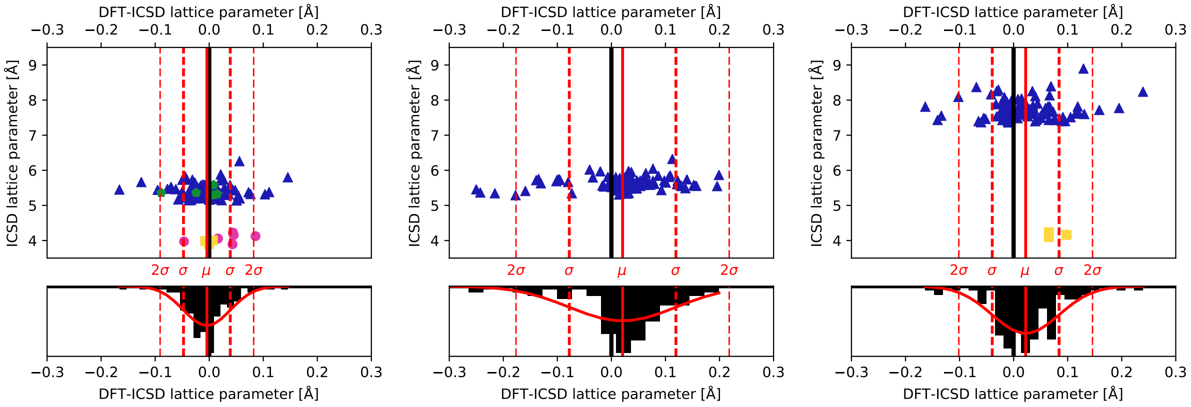


Figure 3.5. Histogram representation of formation energies and band gaps of compounds calculated in this work.



(a) (b) (c)

Figure 3.6. Comparison between DFT and ICSD lattice parameters for 113 compounds: (a) lattice parameter a, (b) lattice parameter b, and (c) lattice parameter c. In the top panels, the horizontal axes measure the difference between the computed and experimental lattice parameters while the vertical axes are the experimental lattice parameters. The lower plots correspond to a histogram of the difference in lattice parameters between DFT and experiment. The solid and dashed red lines indicate the average error, first and second standard deviations between DFT and experiment, respectively.

3.3.3. Supercell Size

To analyze the effect of the size of the supercell used to calculate the oxygen vacancy formation energy of ABO_3 compounds, we compared the vacancy energies using different supercell sizes, and against experiment.[94,99–103] We see good agreement between our high-throughput approach and data from the literature (see Figure 3.4). Previously, Curnan and Kitchin[98] have showed that oxygen vacancy formation energy trend is largely unaffected by the supercell size for $LaBO_3$ and $SrBO_3$ (B = Sc-Cu).

3.3.4. Band Gap

The band gaps were calculated with GGA-PBE, with U-values for some 3d-transition metals and actinides. GGA-PBE tends to underestimate the band gaps of semiconductors[115,116] meaning that band gap values presented in this work have to be taken as lower bound and are useful to identify insulators. Different, much more expensive, calculations, such as hybrid functionals or quasiparticle calculations (G_0W_0, GW_0 and GW), can be done to compute band gap values more accurately.[115]

3.3.5. Magnetism

Several perovskites are experimentally observed to have complex magnetic structures, e.g., antiferromagnetic order.[94] However, only ferromagnetic configurations are calculated in the present study. Stevanovic *et al.*[95], who calculated ternary compounds with up to ten different relative spin orientations, showed that the computational error associated with the wrong magnetic ordering is of the order of 0.01 to 0.02 eV/atom which is not significant for the present study.

3.3.6. Comparison with Experimentally Observed Perovskites

Of the 5,329 different compositions that were calculated, 395 are predicted to be thermodynamically stable by density functional theory. Out of those, 165 are reported in the literature. As a result, 230 new compounds are predicted to be DFT stable but not yet experimentally reported. This set of compounds represents a wide range of predictions amenable for materials synthesis.

The stability values of the 223 compounds that are experimentally reported in the literature are plotted in Figure 3.7. The plot shows that a large number of these experimentally reported compounds are stable according to our DFT T = 0 K calculations. However, the remainder of the phases are above the convex hull, and hence metastable (or unstable). The results of Figure 3.7 shows the measure of metastability in term of convex hull distance: there is rapid decay of the number of synthesized compounds as the convex hull distance increases, reaching almost 0 at a hull distance of 0.1 eV/atom. This 0.1 eV/atom metric for metastability is consistent with the results from another recent high-throughput study of metastability by Sun *et al.*[117]

Nine compounds reported in the literature are seen with a stability above 0.5 eV/atom. All these compounds contain rare earth elements, which are difficult to treat accurately with DFT because of the complexities associated with f-electron systems. In our high-throughput study, f-electrons are not included in the valence electrons of the pseudopotentials used, and therefore the DFT calculations of rare-earth-containing perovskites could have physical errors associated with the approximations made in the DFT calculations. For a more detailed discussion about f-electrons and frozen-core potentials, we refer the reader to Kirklin et al.[62] Error can also come from erroneous experimental characterization and/or classification.

3.4. Usage Notes

We suggest using the data as it is in the spreadsheet. If one chooses to access the data from OQMD via qmpy, we note that the OQMD is a constantly-growing database. Indeed, as a result of compounds being constantly calculated and added to the database,

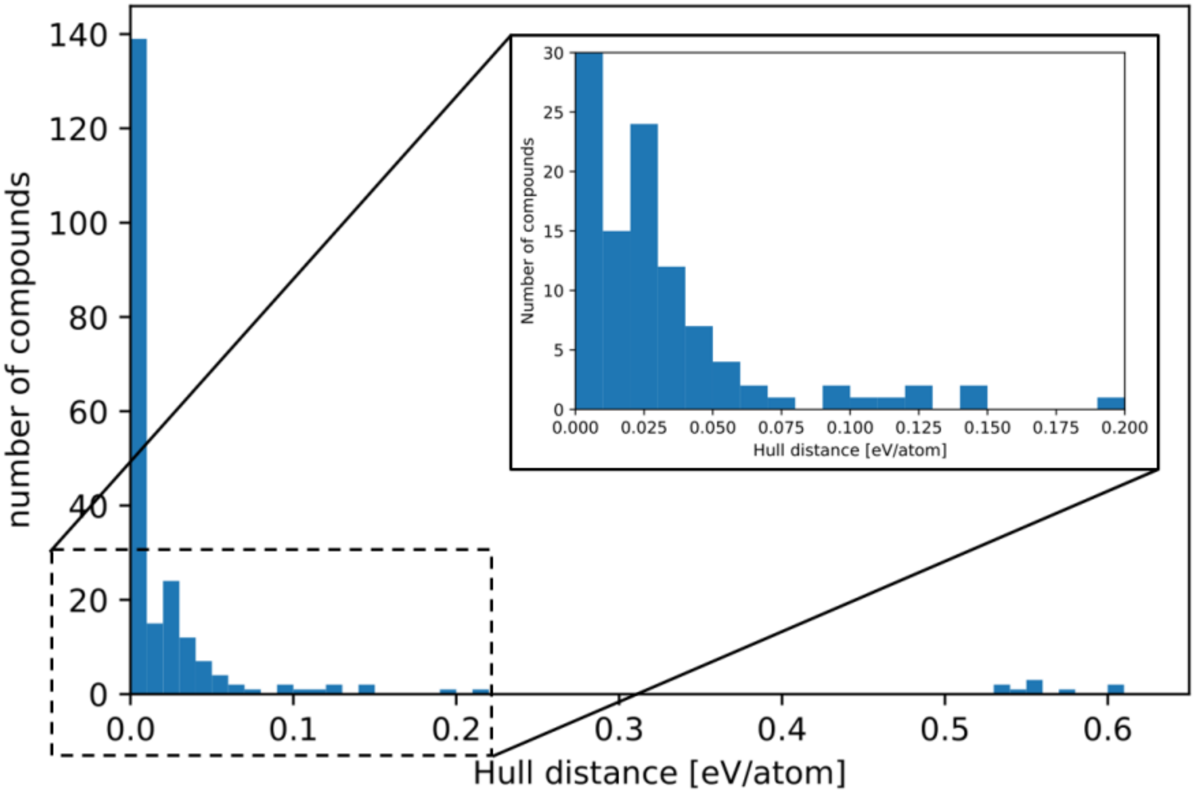


Figure 3.7. Histogram of the DFT stability of 223 ABO_3 perovskite compounds reported in the literature. The inset shows the rapid decay of stable compounds as a function of stability.

the stability of the already-present compounds can change: adding new stable compounds may change the predicted stability of a perovskite.

CHAPTER 4

Computational Screening of Perovskites for Water Splitting Applications

The use of hydrogen as fuel is a promising avenue to aid in the reduction of greenhouse effect gases released in the atmosphere. In this work, we present a high-throughput density functional theory (HT-DFT) study of 5,329 cubic and distorted perovskites ABO_3 compounds to screen for thermodynamically favorable two-step thermochemical water splitting (TWS) materials. From a dataset of more than 11,000 calculations, we screened materials based on: (a) thermodynamic stability, and (b) oxygen vacancy formation energy that allow favorable TWS. From our screening strategy, we identify 139 materials as potential new candidates for TWS application. Several of these compounds, such as $CeCoO_3$ and $BiVO_3$, have not been experimentally explored yet for TWS and present promising avenues for further research. We show that taking into consideration all phases present in the A-B-O ternary phase, as opposed to only calculating the formation energy of a compound, is crucial to assess correctly the stability of a compound as it reduces the number of potential candidates from 5,329 to 383. Finally, our large dataset of compounds containing stabilites, oxidation states and ionic sizes allowed us to revisit the structural maps for perovskites by showing stable and unstable compounds simultaneously.

4.1. Introduction

Hydrogen used as fuel presents several advantages. It can be used in fuel cells for stationary as well as mobile applications and the product of its combustion is water, making it potentially a green alternative to fossil fuels.[118–121] Currently, hydrogen is mainly produced by steam reforming of natural gas[122,123] which decreases the environmental advantage of its carbon free combustion. Thus, exploring ways to produce hydrogen in a sustainable and carbon-neutral fashion, is important for its widespread adoption as a fuel.

Among different carbon-free routes to produce hydrogen, such as photoelectrochemical (PEC) and electrolytic processes, solar driven thermochemical water splitting cycles, i.e. the use of solar thermal energy to drive a set of chemical reactions, have the advantage of using the entire solar spectrum, thus leading to higher theoretical efficiencies.[4,8,10,11,13,15,124] Specifically, hydrogen can be produced by splitting the water molecule in a two-step thermochemical cycle as follows:

$$\mathrm{MO}_x \rightarrow \mathrm{MO}_{x-\delta} + \frac{\delta}{2}\mathrm{O}_2 \tag{4.1}$$

$$\mathrm{MO}_{x-\delta} + \delta\mathrm{H}_2\mathrm{O} \rightarrow \mathrm{MO}_x + \delta\mathrm{H}_2 \tag{4.2}$$

where M is a metal, MO_x its corresponding metal oxide, and δ is the degree of oxygen off-stoichiometry, i.e. the amount of oxygen loss when reducing the metal oxide which is a function of temperature, pressure and the metal oxide. Reaction 4.1 is the thermal reduction (TR) step occurring typically around 2000 K and Reaction 4.2 is the gas splitting (GS) step occurring at lower temperature (around 1000 K).[11] Using a similar concept, carbon monoxide (CO), a hydrocarbon fuel precursor, can be produced by splitting of the carbon dioxide (CO_2) molecule.[13,29] In both cases, the metal oxide and the associated thermodynamics of equation 4.1 and 4.2 are critical aspects that determine the efficiency of the gas splitting.[125,126]

For practical fuel production, metal oxides are required which increase the efficiency and rates at which the thermal reduction and gas splitting reactions occur. Ideally, the materials should be reduced at as low a temperature as possible to avoid energy loss due to heat transfer and thus increase the overall efficiency of the process.[3,8,11,16] However, as the temperature of reduction is also connected to that of the GS, it cannot be too low as to disallow gas splitting. As phase transformations often negatively impact the kinetics and the cyclability of both reactions, it is desirable to have a metal oxide that can accommodate large amounts of oxygen vacancies without changing structure. Other considerations include structural stability at high temperature, as well as cost and availability of the metal oxides. Thus, the choice of the metal oxide is key in terms of the conditions under which solar thermochemical cycles reactions will take place, as well as the capacity of water splitting.

In the literature, 280 thermochemical water-splitting cycles were reviewed by Abanades *et al.*[3] Out of those, 30 were selected as technologically feasible based on temperature of operation, process complexity, cost and toxicity of materials among other criteria.[3] Cycles used successfully to split water include zinc oxide (ZnO/Zn),[16,17] tin oxide (SnO_2/SnO),[18,19] iron oxide[10] and ceria[20–25] based cycles. Ceria was proposed by Otsuka *et al.*[20] as a candidate for water splitting and was studied thoroughly as it appears to be one of the most promising materials for water splitting applications.[20–25] Indeed, it can accommodate a large quantity of oxygen vacancies without phase transformation.[20–25] As a consequence, fast kinetics of thermal reduction and gas splitting reactions were shown by Abanades *et al.*[21] and ceria was cycled 500 times exhibiting stable oxygen release and hydrogen production after 100 cycles.[23] One major drawback of the ceria based cycle is the high temperature needed during the thermal reduction step.

In the search for improved materials for TWS, ABO_3 perovskites have been recently proposed for water splitting.[32–34,127,128] Many perovskites show a high degree of oxygen off-stoichiometry. Furthermore, the alloying potential of both the A and B cation sites makes a very large compositional space for promising materials. Lanthanum based perovskites, such as Mn^{+4} and Sr^{+2} doped $LaAlO_3$, have been experimentally studied and showed promising results in terms of oxygen release during thermal reduction (up to eight times larger than ceria).[33] Additionally, $La_{1-x}Sr_xMnO_3$ exhibited higher solar-to-fuel efficiency than ceria under lower temperature conditions.[34] Finally, Nalbandian *et al.* tested $La_{1-x}Sr_xMO_3$ (M = Mn, Fe, $x = 0, 0.3, 0.7, 1$) perovskites in a membrane reactor to produce hydrogen in an isothermal and continuous fashion.[32]

The remarkable stability of the perovskite structure with respect to its constituents elements and its ability to tolerate a high degree of oxygen off-stoichiometry, suggests that there could be potentially superior perovskite compounds awaiting discovery.[39] As the number of possible compositions for ABO_3 structures is too large to be completely explored experimentally, we use our dataset of ABO_3 compounds that was calculated in chapter 3 to search efficiently for novel perovskite metal oxides for thermochemical water splitting applications. Today's computational resources enable high-throughput DFT where large numbers of compounds (typically ten or hundred thousands) are calculated to create databases containing energies of numerous materials.[58–61,96,129] Databases can then be searched for materials with desired properties, thus accelerating materials design, as shown successfully for various applications,[63–66,96] most notably a perovskites search for photoelectrochemical water splitting.[130,131] Here, we build on such approaches to screen for thermochemical water splitting perovskites.

We start with the central premise that tuning the thermodynamics of oxide reduction and compound stability is necessary to enable improved material performance of the thermochemical water splitting process. Two filters are used to screen our high-throughput DFT dataset for potential water splitting oxides: 1) Stability: A perovskite compound must be thermodynamically stable in order to be considered as a promising candidate. In this work, stability refers to 0 K and 0 bar DFT stability of ABO_3 with respect to all possible combinations of phases present in the A-B-O ternary phase diagram. In other words, the ABO_3 perovskite should be on the ground-state convex hull, which is defined as the envelope connecting the lowest energy compounds at every composition in the phase space.[132] 2) Oxygen vacancy formation energy: In order to split water, both thermal reduction and gas splitting reactions must be thermodynamically favorable, i.e. have a negative Gibbs free energies. Meredig and Wolverton[125,126] showed that, for typical TR and GS temperatures, these conditions impose limits on the enthalpy and entropy of reduction, creating a window where both reactions are thermodynamically favorable. In the present study, we use these limits on the enthalpy of reduction as a second filter. To avoid false negative classification of compounds and to account for the approximations inherent to the high-throughput character of this study, we use a slightly larger oxygen vacancy formation energy window; we consider compounds that have an oxygen vacancy formation energy between 2.5 and 5.0 eV/O atom to be suitable for TWS.

After calculating 5,329 different compositions in the cubic, rhombohedral, tetragonal and orthorhombic distortions of perovskite, we found 383 stable compounds. Among those, 139 fell within the suitable oxygen vacancy formation energy range for water splitting. In addition, we identified that rare earth elements and 3d-transition metals are prominent on the A- and B-sites of stable perovskites, respectively. The high-throughput database of compounds and energies allowed us to reconsider structural maps for the perovskite phase across a very wide range of possible A- and B-site chemistries.

4.2. Methodology

For all details about the calculation parameters, compounds calculated and properties computed, we refer the reader to the methodology section of chapter 3.

4.3. Results and Discussion

4.3.1. Stability

The results for perovskite stability of all compositions are summarized in Table 4.1. A vast majority (92%) of the compounds have negative formation energy ($\Delta H_f^{\mathrm{ABO_3}}$ as defined in equation 3.2). This result is expected as we are combining electropositive elements, in our case metals and semi-metals, to oxygen, the second highest electronegative element. A negative formation energy only indicates stability of a compound with respect to its constituent elements and is thus a necessary but not sufficient condition for stability. In this work, we consider compounds with a stability ($\Delta H_{\mathrm{stab}}^{\mathrm{ABO_3}}$) lower than 25 meV/atom (approximately kT at room temperature) to be either stable or nearly so. Out of all the compounds that have $\Delta H_f^{\mathrm{ABO_3}} < 0$, only a small fraction (4%) are stable with respect to all the phases present in the A-B-O convex hull. We report 383 stable perovskite compounds which is considerably more than what was reported in a similar recent study.[133] We explain this discrepancy by the fact that, as opposed to Körbel *et al.*, we included lanthanides as potential cations. Indeed, lanthanide-containing perovskites represent approximatively 50% of the compounds we predict to be stable. In Table 4.1, rhombohedral, tetragonal or orthorhombic compounds that are stable and relax to a higher symmetry group (i.e rhombohedral to cubic, tetragonal to cubic, tetragonal to rhombohedral, orthorhombic to tetragonal, orthorhombic to rhombohedral or orthorhombic to cubic) are labeled as stable in their highest symmetry group. The orthorhombic distortion accounts for 84% of all the stable perovskites, which is in accordance with the literature stating that space group 62 is the most common perovskite structure.[134] The difference between negative formation energy and stability shows the importance of having a large and complete database, including theoretical prototypes, available when computing the high-throughput stability of compounds by first-principles calculations. The database is of particular relevance when dealing with compounds that contain elements that have not been thoroughly explored experimentally as missing phases in those systems can lead to inaccurate stability calculations.

Our high-throughput study allows us to obtain a complete picture of the stability of the perovskite compounds in terms of frequency of elements on the A- and B-sites. Figure 3.1 shows the number of stable occurrences for each elements on the A- or B-site. Out of the 73 elements we considered for this study, 44 appear in stable perovskites on the A-site and 57 appear on the B-site. Three elements (B, Mg, Zn) do not appear in any stable ABO_3 perovskites, regardless of the other metal or whether these elements are placed on the A or B sites. Conversely, 30 elements are predicted to form perovskites in both the A- and the B-site. Numerous perovskites containing elements that are not extensively studied experimentally such as lanthanides and actinides are predicted to be stable and present opportunities for new materials discovery. Alkali metals, alkaline earths, rare earths and 3d-transition metals are heavily represented as elements forming stable perovskite structure which is consistent with other work in the literature.[38,110] To see similar behavior among elements from the same group, we plotted the stable perovskites in a network graph by clustering elements by groups (see Figure 4.1). In this plot, each connecting line represents a stable compounds (i.e. 383 tie-lines are drawn). Each disk represents an element that is present in at least one stable perovskite (i.e. 70 disks are displayed). Elements are colored and clustered by groups and their size is proportional to the amount of lines connected to them. The curvature of the lines is always clockwise which gives an information on which elements is on the A- and the B-site. For instance, the lines connecting the rare earth and the 3d transition metals are all convex, indicating that the rare earths are on the A-sites and the 3d transition metals are on the B-sites. We observe a high occurrence of alkali metals, alkaline earths and rare earths on the A-sites and 3d-transition metals and 4d-transition metals on the B-sites. This picture is in agreement with similar studies found in literature.[39]

We find a total of 383 stable perovskites, shown in Figure 4.2. We observe a cluster of compounds containing rare earth on the A-site and transition metals on the B-site, which is likely due to the multiple possible oxidation states for those elements, allowing them to be in the +1/+5, +2/+4 or +3/+3 A/B cation oxidation state configuration and form charge neutral compounds with numerous other elements. This result is in agreement with the observation that the B-site atom is almost always smaller the A-site atoms in the perovskite structure as it is embedded in a rigid octahedron of oxygen.[135–137]

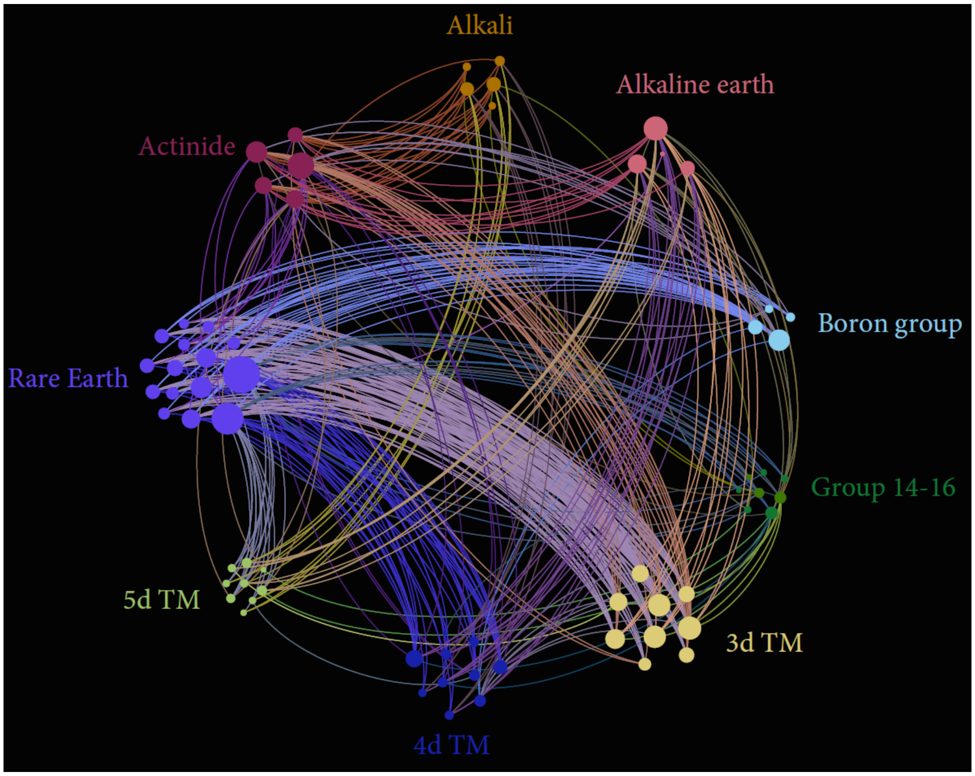


Figure 4.1. Network graph of stable perovskites in terms of cation frequency. Each disk represents an element that can be on the A- or B-site, their size is proportional to the number of lines connected to them and their color and positioning helps clustering elements from the same group. Each line corresponds to a stable compound, their color is a mixing of both disks that they connect. The curvature of the lines is always clockwise. TM stands for transition metals.

Additionally, we see very few compounds containing transition metals (rows labeled 3d, 4d and 5d in Figure 4.2) on the A-site, probably for a similar reason: transition metals are generally smaller than the other elements and thus preferentially tend to occupy the B-site. Conversely, most of the perovskites predicted to be stable in the undistorted cubic geometry contain bigger elements such as alkali and alkaline-earth metals (Li, Be, Na, Mg, K, Ca, Rb, Sr, Cs and Ba) on the A-site. It is worth noting that the high occurrence of distorted orthorhombic perovskites showcases their 0 K ground-state. However, perovskites often exhibit a phase transformation between low and high temperature, which is likely to happen at temperatures required for TWS.[39] With so few rhombohedral and tetragonal perovskites, no obvious trends emerge in their constituting elements.

Table 4.1. Breakdown of calculated stable perovskites by the type of structural distortion favored. $\Delta H_f^{\mathrm{ABO_3}}$ corresponds to the formation energy with respect to the constituent elements of the perovskite and $\Delta H_{\mathrm{stab}}^{\mathrm{ABO_3}}$ is the stability of a compound with respect to all other phases present in the A-B-O phase diagram.

Distortion	Number calculated	$\Delta H_f^{\mathrm{ABO_3}}$ <0 meV/at.	$\Delta H_{\mathrm{stab}}^{\mathrm{ABO_3}}$ <25 meV/at.
Cubic	5329	4778	41
Rhombohedral	2162	2107	15
Tetragonal	2162	1910	5
Orthorhombic	2162	2041	322
Total	**11,815**	10,836	**383**

4.3.2. Discovery of New, Stable Perovskite Compounds

We wish to analyze which of our list of stable perovskites are already known experimentally, and which represent new, predicted stable compounds. In order to compare our list of 383 predicted stable perovskites with those found in the experimental literature, we have compiled a list of experimentally reported perovskites. Four perovskite review papers (Roth,[110] Giaquinta *et al.*,[138] Li *et al.*[139] and Zhang *et al.*[140]) were aggregated to form a list of 251 experimentally observed perovskites. Out of those 251 perovskites, we predict 170 compounds to be stable. DFT predicts accurately the stability of all the well-known and studied perovskites such as $BaTiO_3$, $SrTiO_3$, $LaAlO_3$ and $CaTiO_3$. Out of the 383 compounds we predict to be stable, 213 of them are not reported in the literature and thus, represent great potential for discovery of new compounds, even outside of the thermochemical water splitting application. For instance, $BiVO_3$, $CeCoO_3$, $CeAgO_3$ and $LaAgO_3$, $YbMoO_3$ and $LiIrO_3$ are all perovskites predicted to be stable here but, to our knowledge, have not been synthesized yet and hence, are amenable for experimental testing. An extended list of such compounds is given in Table 4.3. We find 81 compounds that are predicted to be unstable but were found in at least one review paper. Among those, we found 26 compounds that have different ABO_3 stable phases in the OQMD, such as ilmenite ($R\bar{3}$, 148) or the hexagonal distortion ($P6_3cm$, 185). Thus, we suspect that those compounds are reported in the literature in a distortion that was not considered in this high-throughput study. The other 55 compounds are unstable and are predicted to decompose in a linear combination of more stable unary, binary or ternary phases. Finally, it is worth noting that among our list of 213 compounds predicted to be stable, 97 contain actinides used in this study (Ac, Th, Pa, U, Np and Pu) or unstable elements (Tc and Pm). Other elements that are not often encountered in the literature, such as Europium and Ytterbium are also present in 29 and 26 compounds, respectively. The high occurrence of such compounds might be attributed to the lack of competing phases present in the database during the stability calculation, potentially mislabeling them as stable. However, the prediction of a stable phase at the ABO_3 composition indicates that either the perovskites structure is stable or some other, undiscovered, phase(s) must be present at that composition to remove the perovskite from the convex hull.

4.3.3. Oxidation States

For this high-throughput study, we performed calculations for all combinations of metals in A and B perovskite sites, regardless of metal oxidation state. This approach was chosen for several reasons: 1) We did not want to rely on chemical intuition for oxidation states, as it can sometimes fail, which was shown in previous computational studies where ground-state structures of compounds containing elements from the same group have different crystal structures.[141] 2) We expect the majority of the charge unbalanced compounds to be unstable, giving us a useful test of our stability filter. 3) Nine elements present in our study (Tc, Os, Pt, Au, Pm, Ac, Pa, Np and Pu) do not have bond valence parameters, complicating the determination of their oxidation state.

Table 4.2. Oxidation state breakdown for all the 5,329 calculated perovskite compounds. Oxidation states were calculated using bond valence parameters[104] as implemented in pymatgen.[105]

Oxidation State	All compounds		Stable compounds	
A-site: +1, B-site: +5	222		19	
A-site: +2, B-site: +4	884		93	
A-site: +3, B-site: +3	935		143	
A-site: +4, B-site: +2	146		1	
A-site: +5, B-site: +1	28		-	
Total charge balanced		**2215**		**256**
Charge unbalanced	1881		13	
Total charge unbalanced		**1881**		**13**
Contains elements not in BVM	1233		114	
Total unknown charge		**1233**		**114**
Total		**5329**		**383**

The different oxidation state configurations for all the compounds are presented in Table 4.2. Less than half the compounds (2,215 out of 5,329) are predicted to be charge balanced using the bond valence method. By removing compounds containing the 9 elements without BVM parameters, we report 1,233 compounds $((2 * 73 - 9) * 9)$ that have unknown oxidation states. As expected, the vast majority of the charge unbalanced compounds are predicted to be unstable. However, we note that 13 compounds ($AsKO_3$, $AsRbO_3$, $BaBiO_3$, $BeEuO_3$, $CaBiO_3$, $CdBiO_3$, $KTeO_3$, KUO_3, $SiEuO_3$, $SmScO_3$, $TeEuO_3$, $TeHgO_3$ and $TeSrO_3$) are calculated to be stable despite being predicted as charge unbalanced by the bond valence method. Additionally, 114 compounds which have oxidation states that do not have BVM parameters turn out to be stable. These two results emphasis the merit of considering all 5,329 compounds for our high-throughput study, as we would have missed those 127 compounds (about one third of all the stable compounds) by doing a screening before performing the DFT calculations. We see that the majority of compounds have both cations in the +3 or +2 on the A-site and +4 B-site probably because those oxidations states are the most common in the periodic table. On the other hand, fewer compounds are in the +5 and +1 configuration, as those oxidation states are less common. There are more +1/+5 and +2/+4 compounds than +5/+1 and +4/+2 respectively, which is consistent with the knowledge than the A-atom is most often the biggest of the two cations.[39]

4.3.4. Perovskite Structure Maps

With the large database of perovskites and their stability at hand, it is possible to draw structure maps of perovskite stability similarly to what has been done in other previous experimental and theoretical studies.[41,110,138,139,142–144] Structure maps are tools designed to help classify the structure of stable compounds as a function of the radii of their constituent elements. They have been previously used in the literature to predict the stability of a variety of different compounds, including perovskites.[110,138,139,142,143] In our study, we used ionic radii, as tabulated by Seshadri and Basu,[108] to create a radius of A vs B atom (r_A vs r_B) structure map (Figure 4.3). Figure 4.3 shows that 95% of the stable perovskites lie in the upper left region, i.e. where $r_A > r_B$. This observation is in agreement with what is reported in the literature where the same r_A vs r_B map was used with experimentally observed perovskites.[39,138] We also see that, by removing the charge unbalanced compounds, we are mostly removing the perovskites that have large B-atoms. Due to the difference in coordination number between the A- and B-atom (12 and 6 respectively), and thus the difference in radius for the ABO_3 vs BAO_3 compounds, Figure 4.3 is not symmetric with respect to the $x = y$ line. Although the stable compounds are highly clustered in Figure 4.3, some outliers indicates that the simple purely geometrical argument is not always sufficient to describe the stability of perovskites. This result is consistent with a recent density functional theory study of perovskites.[144] Also, the region with clustered stable perovskites contains a large number of unstable compounds. So, while the map shows a high degree of clustering, it does not show a high degree of separation between stable and unstable compounds. A highly predictive structure map needs to have both clustering and separation of data.

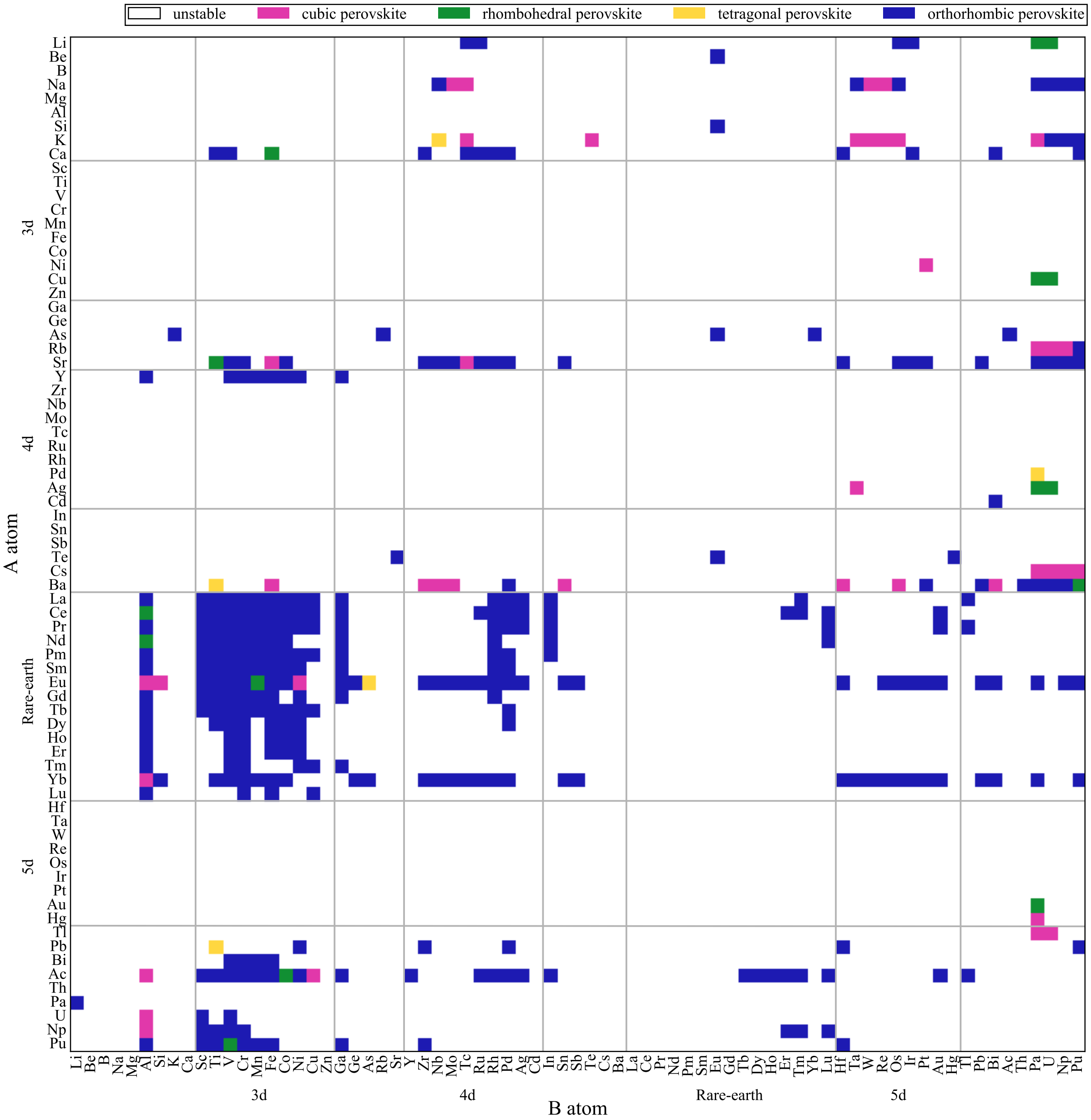


Figure 4.2. High-throughput DFT stability map of perovksites compounds. Space of all the 5,329 stoichiometries considered, and the 389 perovskites that are predicted to be stable with respect to all the phases in the A-B-O phase diagram. The atoms on the x- and y-axis are ordered by atomic number. The colors show which distortions are stable.

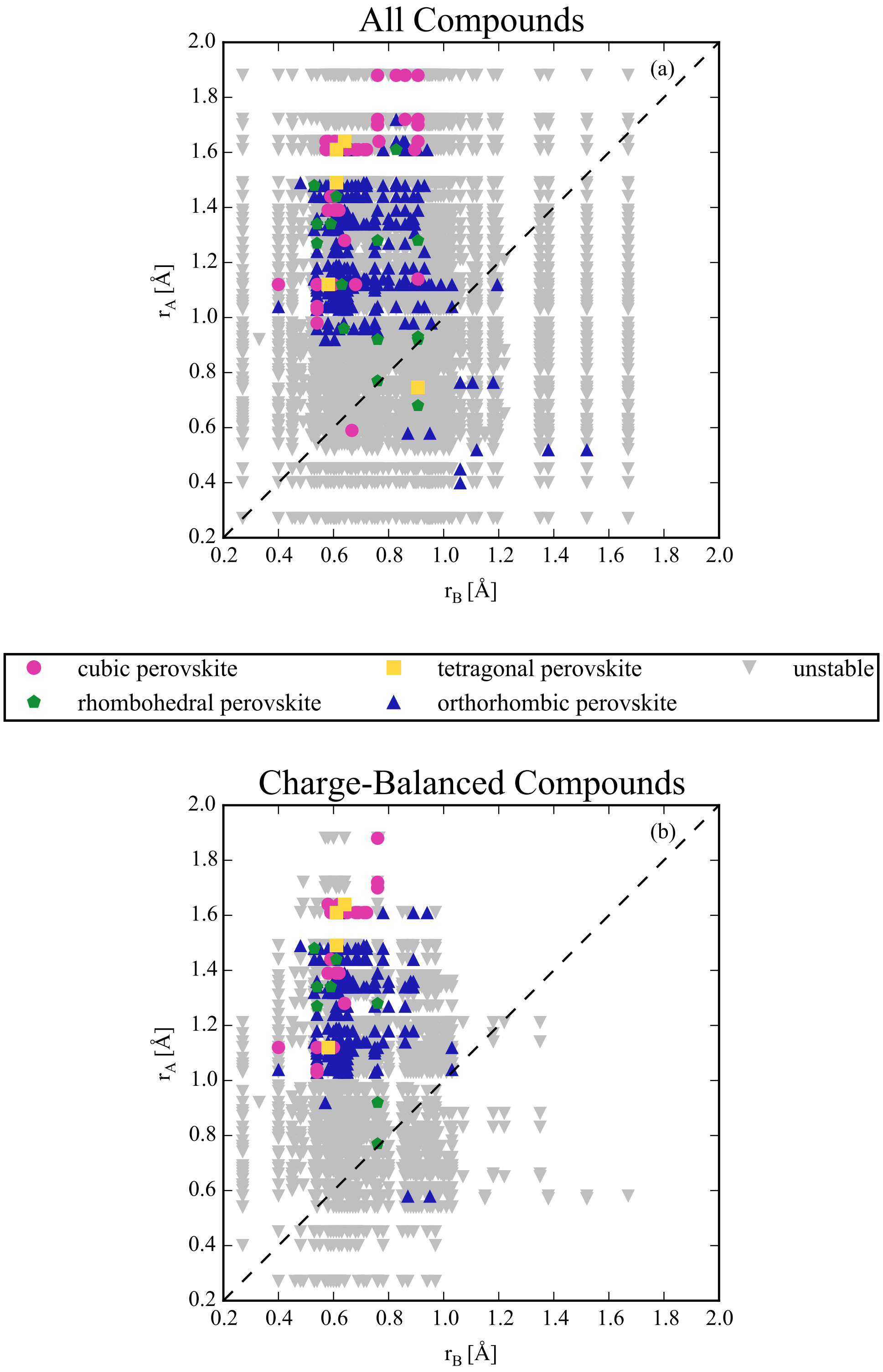

All Compounds
(a)
r_A [Å]
r_B [Å]
2.0
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
2.0
● cubic perovskite
■ tetragonal perovskite
▼ unstable
⬟ rhombohedral perovskite
▲ orthorhombic perovskite
Charge-Balanced Compounds
(b)
r_A [Å]
r_B [Å]
2.0
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
2.0

Figure 4.3. Perovskite structure maps of radii of A- and B-cations (r_A vs r_B). Colored symbols are stable perovskites and grayed out symbols are unstable. After calculating the oxidation state of both cations for each compounds using the bond valence method,[104] Shannon radii were used for each element.[106,107] The dashed line represents the $x = y$ line. (a) Plot containing all compounds (5,329 compositions) and (b) Plot containing only charge balanced compounds (2215 compositions).

It is also interesting to look at the tolerance factor, defined by equation 2.5, distribution of stable and unstable compounds. Figure 4.4 represents the frequency at which stable and unstable perovskites with different tolerance factors appear in our dataset. Most of the stable perovskites have a tolerance factor comprised between 0.8 and 1.1 which is consistent with other studies that used the tolerance factor as a way to separate stoichiometries that would form perovskite.[39,41] However, we observe that a large number of compounds with a tolerance factor within this window are calculated to be unstable. Approximately, 750 charge balanced perovskites with a tolerance factor between 0.9 and 1 are predicted to be unstable. This indicates that, when taking a random composition, geometrical and charge neutrality arguments are not always sufficient to describe the stability of a perovskite.

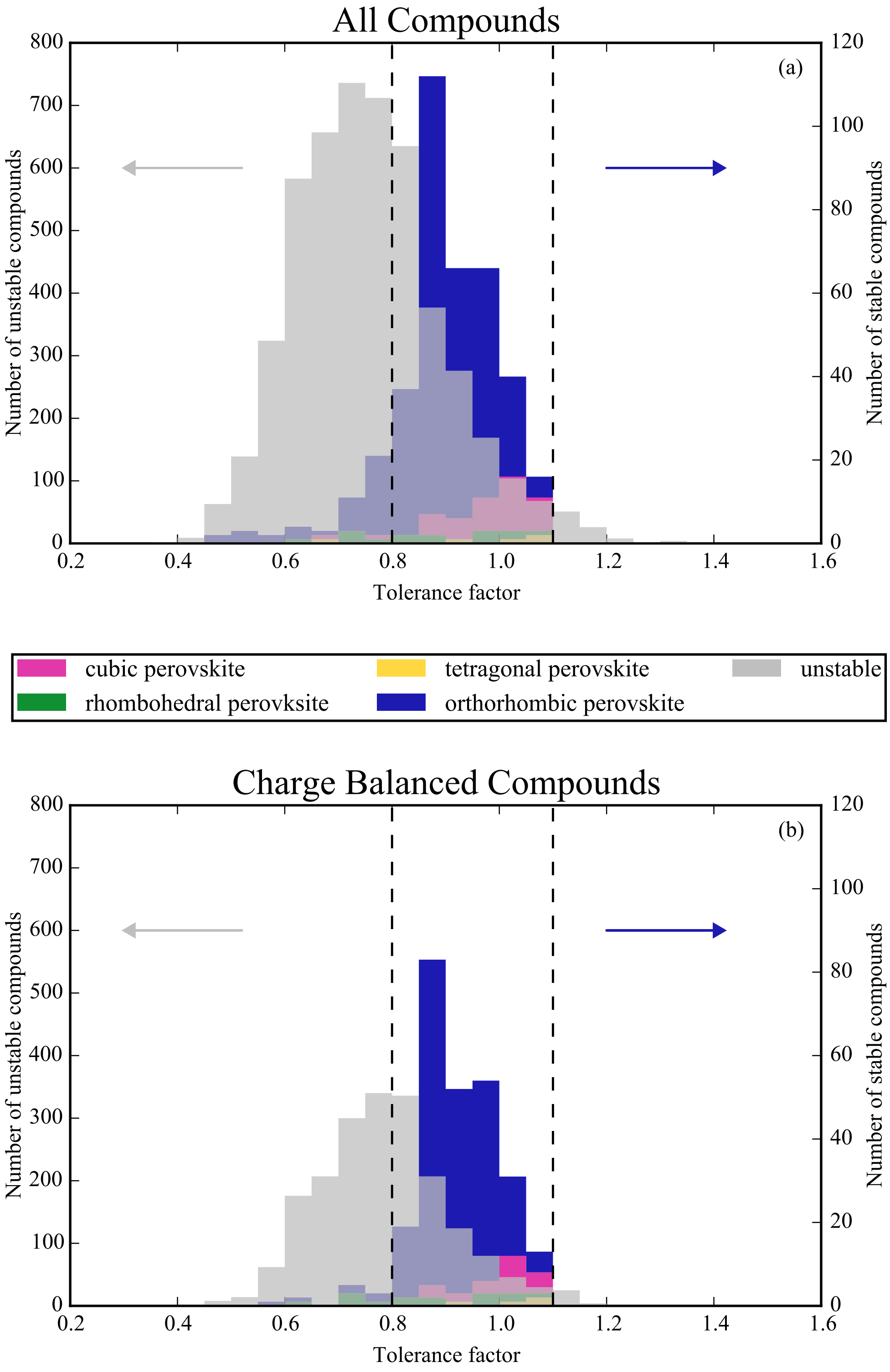

All Compounds
(a)
Number of unstable compounds
Number of stable compounds
Tolerance factor
0
100
200
300
400
500
600
700
800
0
20
40
60
80
100
120
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
cubic perovskite
tetragonal perovskite
unstable
rhombohedral perovksite
orthorhombic perovskite
Charge Balanced Compounds
(b)
Number of unstable compounds
Number of stable compounds
Tolerance factor
0
100
200
300
400
500
600
700
800
0
20
40
60
80
100
120
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6

Figure 4.4. Frequency of stable perovskites in function of their tolerance factor. Shannon radii were used for to compute the tolerance factors of each compound.[106,107] The y-axis is normalized by the number of stable perovskites (389) for the stable compounds and unstable compounds (4940) for the unstable compounds. (a) Plot containing all compounds (5,329 compositions) and (b) Plot containing only charge balanced compounds (2215 compositions).

4.3.5. Selection of Novel Materials for Water Splitting

By using the stability and the oxygen vacancy formation energy filters, we can now screen for novel materials for water splitting. Figure 4.5 (a) shows the extent of the present work by plotting the calculated compounds in a single plot where the x-axis corresponds to the stability and the y-axis the oxygen vacancy formation energy. For the 2,162 cases where distortions were calculated, we plot the distortion with the lowest energy; for the cases where the distortions were not calculated, we plot the energy of the cubic phase. We clearly see the 0.5 eV/atom threshold below which structural distortions were calculated. The distorted points shown above the 0.5 eV/atom threshold represent the compounds that were randomly selected to assess the impact of the distortions on the stability of perovskites. We note that, due to the use of a cubic supercell to calculate the oxygen vacancy formation energies, a couple of distorted compounds have negative oxygen formation energy and are stable. Figure 4.5 (b) shows the region where stable perovskites are found. Out of the 383 compounds that are calculated to be stable, 139 (12 cubic, 8 rhombohedral and 119 orthorhombic) fall in the target window i.e. have an oxygen vacancy formation energy between 2.5 and 5 eV/O atom. Ceria stability was calculated through the OQMD and added to the plot along with oxygen vacancy formation energies measured experimentally[145] and calculated from first-principles using different techniques.[146,147] It falls within our target windows, giving us confidence in our filtering method. All compounds passing both filters are shown in Table 4.3. Owing to their ability to give or receive electrons, and thus enforce the charge neutrality of a compound, elements with multiple possible oxidation states such as transition metals and rare earth elements are heavily represented in this category. Many of the lanthanum based perovskites that were mixed together ($LaMnO_3$, $LaCrO_3$ and $LaFeO_3$)[32–34,127,128] and used for water splitting are also in Table 4.3. Finally, ceria-containing compounds appear frequently in the list of potential candidate.

Perovskites containing earth-abundant elements such as $CaVO_3$, $SrVO_3$ and $SrSnO_3$, compounds not reported in the literature such as $BiVO_3$ and $CeCoO_3$ or compounds that are close to ceria on Figure 4.5 (b) such as $EuGeO_3$ and $EuSnO_3$, were all predicted to be good candidates for water splitting application and might be of special interest to the reader. On the other hand, some perovskites present in Table 4.3 contains uncommon elements such as europium and actinium. Even if such compounds are not likely to be used experimentally, we chose to keep them in our high-throughput study as it might open new amenities for completely different compositions.

Other descriptors could be used to narrow down the number of candidates. For instance, previous studies showed that large and positive entropy is a key parameter to insure thermodynamically viable reactions.[125,126] Additionally, reaction kinetics is also critical for practical fuel production, particularly for the low-temperature water splitting step. Even though surface reactions are often complex, especially for ceria,[148,149] one could imagine kinetic arguments to design additional descriptors. For example, oxygen diffusivity could play a role in the reaction kinetics and could be used as a filter to further enhance our screening capabilities.

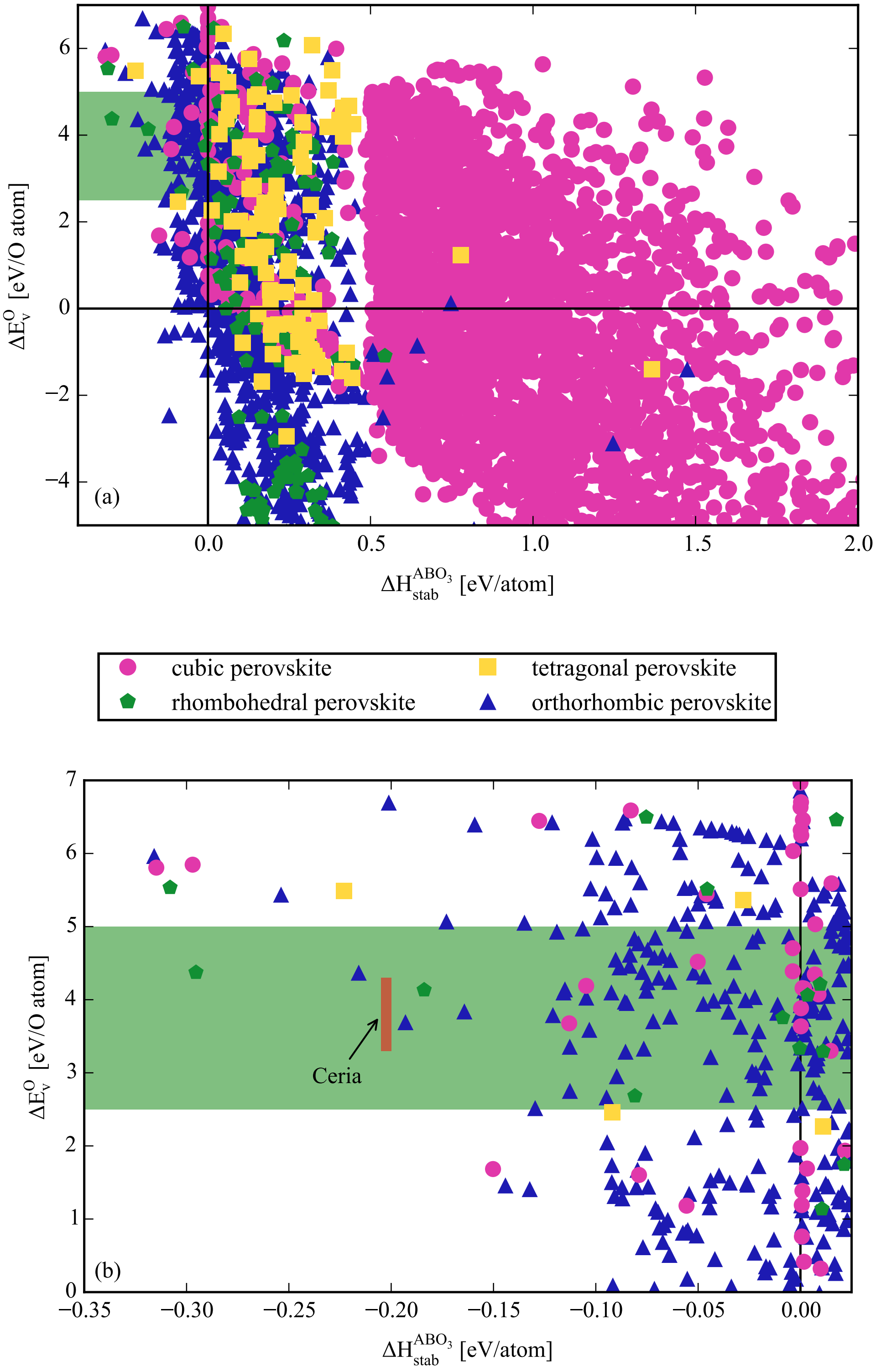

(a)
ΔE_v^O [eV/O atom]
6
4
2
0
−2
−4
0.0
0.5
1.0
1.5
2.0
$\Delta H_{stab}^{ABO_3}$ [eV/atom]
cubic perovskite
tetragonal perovskite
rhombohedral perovskite
orthorhombic perovskite
(b)
ΔE_v^O [eV/O atom]
7
6
5
4
3
2
1
0
−0.35
−0.30
−0.25
−0.20
−0.15
−0.10
−0.05
0.00
$\Delta H_{stab}^{ABO_3}$ [eV/atom]
Ceria

Figure 4.5. Perovskite oxygen vacancy formation energy (ΔE_v^{O}) plotted against stability ($\Delta H_{\mathrm{stab}}^{\mathrm{ABO_3}}$). The green windows represents the oxygen vacancy formation energy target. (a) The stabilities of distortions with the lowest energy are plotted. For the 3167 (5329-2162) compounds where we did not calculate distortions, the cubic stability is plotted. (b) Blowup of the stable phases region. Experimental and calculated oxygen vacancy formation energies of ceria are taken from Chiang *et al.*,[145] Yang *et al.*[146] and Murgida *et al.*[147]

Table 4.3. List of perovskites, ordered by ascending vacancy formation energy, for water splitting application. All these compounds pass the stability ($\Delta H_{\mathrm{stab}}^{\mathrm{ABO_3}} < 0.025$ eV/atom) and oxygen vacancy formation energy ($2.5 < \Delta E_v^{\mathrm{O}} < 5$ eV/O atom) screens. Compounds in bold are not reported in any of the review papers used for the literature survey.[110,138–140] Oxidation states were calculated using bond valence parameters[104] as implemented in pymatgen.[105]

Formula	Stable Distortion	$\Delta H_{\mathrm{stab}}^{\mathrm{ABO_3}}$ [eV/atom]	ΔE_v^{O} [eV/O atom]	Ox. N.$_A$	Ox. N.$_B$
$TbCoO_3$	orthorhombic	0.001	2.517	3	3
$EuOsO_3$	**orthorhombic**	**-0.130**	**2.517**	**unknown**	
$BiMnO_3$	orthorhombic	0.019	2.550	3	3
$BiFeO_3$	orthorhombic	0.012	2.593	3	3
$NaNpO_3$	**orthorhombic**	**-0.095**	**2.664**	**unknown**	
$AuPaO_3$	**rhombohedral**	**-0.081**	**2.683**	**unknown**	
$EuSbO_3$	**orthorhombic**	**-0.036**	**2.699**	**3**	**3**
$LuFeO_3$	orthorhombic	0.009	2.732	3	3
$EuRuO_3$	**orthorhombic**	**-0.113**	**2.753**	**2**	**4**
$YbPuO_3$	**orthorhombic**	**0.024**	**2.789**	**unknown**	
$AcRuO_3$	**orthorhombic**	**0.006**	**2.793**	**unknown**	
$SmCoO_3$	orthorhombic	-0.018	2.937	3	3
$YbTcO_3$	**orthorhombic**	**0.003**	**2.953**	**unknown**	
$LaCoO_3$	orthorhombic	-0.023	2.970	3	3
$PmCoO_3$	**orthorhombic**	**-0.022**	**2.996**	**unknown**	
$NaOsO_3$	**orthorhombic**	**-0.024**	**3.020**	**unknown**	
$LiTcO_3$	**orthorhombic**	**0.006**	**3.047**	**unknown**	

Table 4.3 – continued

Formula	Stable Distortion	$\Delta H_{\text{stab}}^{\text{ABO}_3}$	ΔE_v^{O}	Ox. N.$_A$	Ox. N.$_B$
		[eV/atom]	[eV/O atom]		
$HoFeO_3$	orthorhombic	0.008	3.100	3	3
$NdCoO_3$	orthorhombic	-0.001	3.130	3	3
$SrRuO_3$	orthorhombic	-0.017	3.142	2	4
$YFeO_3$	orthorhombic	-0.113	3.196	3	3
$DyFeO_3$	orthorhombic	0.008	3.199	3	3
$CeCoO_3$	**orthorhombic**	**-0.044**	**3.216**	**3**	**3**
$TbFeO_3$	orthorhombic	0.008	3.270	3	3
$EuReO_3$	**orthorhombic**	**-0.085**	**3.289**	**2**	**4**
$PmInO_3$	**orthorhombic**	**0.023**	**3.292**	**unknown**	
$CuUO_3$	**rhombohedral**	**0.011**	**3.294**	**1**	**5**
$KOsO_3$	**cubic**	**0.015**	**3.300**	**unknown**	
$SrOsO_3$	**orthorhombic**	**-0.077**	**3.331**	**unknown**	
$AcCoO_3$	**rhombohedral**	**0.000**	**3.336**	**unknown**	
$BiVO_3$	**orthorhombic**	**0.000**	**3.346**	**3**	**3**
$PrInO_3$	**orthorhombic**	**0.005**	**3.349**	**3**	**3**
$YbSnO_3$	**orthorhombic**	**0.023**	**3.357**	**2**	**4**
$GdFeO_3$	orthorhombic	-0.022	3.390	3	3
$NdInO_3$	orthorhombic	0.013	3.400	3	3
$LaInO_3$	orthorhombic	0.018	3.403	3	3
$CaTcO_3$	**orthorhombic**	**0.017**	**3.425**	**unknown**	
$YMnO_3$	orthorhombic	0.011	3.474	3	3
$TbMnO_3$	orthorhombic	0.002	3.526	3	3
$CeInO_3$	**orthorhombic**	**-0.004**	**3.536**	**3**	**3**
$CaVO_3$	orthorhombic	-0.020	3.569	2	4
$EuPuO_3$	**orthorhombic**	**-0.103**	**3.589**	**unknown**	
$GdMnO_3$	orthorhombic	0.002	3.606	3	3
$SmFeO_3$	orthorhombic	-0.029	3.630	3	3
$NaTcO_3$	**cubic**	**0.000**	**3.635**	**unknown**	
$EuNpO_3$	**orthorhombic**	**-0.001**	**3.640**	**unknown**	
$PuZrO_3$	**orthorhombic**	**0.021**	**3.640**	**unknown**	
$BaOsO_3$	**cubic**	**-0.113**	**3.677**	**unknown**	
$EuSnO_3$	**orthorhombic**	**-0.193**	**3.692**	**2**	**4**
$PmFeO_3$	**orthorhombic**	**-0.072**	**3.706**	**unknown**	
$SmMnO_3$	orthorhombic	0.001	3.736	2	4
$AgUO_3$	**rhombohedral**	**-0.009**	**3.754**	**1**	**5**
$PmMnO_3$	**orthorhombic**	**-0.064**	**3.769**	**unknown**	
$SrPuO_3$	orthorhombic	-0.121	3.789	unknown	

Table 4.3 – continued

Formula	Stable Distortion	$\Delta H^{\mathrm{ABO_3}}_{\mathrm{stab}}$	$\Delta E^{\mathrm{O}}_{v}$	Ox. N.$_A$	Ox. N.$_B$
		[eV/atom]	[eV/O atom]		
$SrVO_3$	orthorhombic	-0.001	3.792	2	4
$TmVO_3$	orthorhombic	-0.011	3.822	2	4
$YbVO_3$	orthorhombic	-0.105	3.829	2	4
$NdFeO_3$	orthorhombic	-0.035	3.836	3	3
$EuTcO_3$	**orthorhombic**	**-0.164**	**3.840**	**unknown**	
$BiCrO_3$	**orthorhombic**	**0.012**	**3.849**	**3**	**3**
$NdMnO_3$	orthorhombic	0.005	3.849	3	3
$KTcO_3$	**cubic**	**0.000**	**3.883**	**unknown**	
$ErVO_3$	orthorhombic	0.005	3.887	3	3
$PrMnO_3$	orthorhombic	0.002	3.889	3	3
$LaMnO_3$	orthorhombic	0.000	3.927	3	3
$LuCrO_3$	orthorhombic	-0.011	3.932	3	3
$HoVO_3$	**orthorhombic**	**0.006**	**3.932**	**2**	**4**
$SrSnO_3$	orthorhombic	-0.047	3.943	2	4
$CeMnO_3$	orthorhombic	-0.085	3.955	3	3
$PrFeO_3$	orthorhombic	-0.086	3.974	3	3
$PuHfO_3$	**orthorhombic**	**-0.042**	**3.987**	**unknown**	
YVO_3	orthorhombic	-0.037	3.990	3	3
$AcInO_3$	**orthorhombic**	**-0.027**	**3.994**	**unknown**	
$DyVO_3$	orthorhombic	0.005	3.995	2	4
$YbGeO_3$	**orthorhombic**	**0.001**	**4.029**	**2**	**4**
$TmCrO_3$	orthorhombic	-0.022	4.037	3	3
$LaFeO_3$	orthorhombic	0.007	4.043	3	3
$TbVO_3$	orthorhombic	0.004	4.045	2	4
$YbWO_3$	**orthorhombic**	**-0.075**	**4.046**	**2**	**4**
$UScO_3$	**orthorhombic**	**0.013**	**4.052**	**3**	**3**
$LiUO_3$	**rhombohedral**	**0.003**	**4.062**	**1**	**5**
$CsUO_3$	**cubic**	**0.009**	**4.073**	**1**	**5**
$CeFeO_3$	orthorhombic	-0.092	4.100	3	3
$AcMnO_3$	**orthorhombic**	**-0.116**	**4.100**	**unknown**	
$GdVO_3$	orthorhombic	0.005	4.112	3	3
$EuVO_3$	**orthorhombic**	**-0.075**	**4.123**	**2**	**4**
$ErCrO_3$	orthorhombic	0.002	4.126	3	3
$CuPaO_3$	**rhombohedral**	**-0.184**	**4.134**	**unknown**	
$AcFeO_3$	**orthorhombic**	**-0.115**	**4.141**	**unknown**	
$SrTcO_3$	**cubic**	**0.002**	**4.154**	**unknown**	
$BaSnO_3$	cubic	0.001	4.158	2	4

Table 4.3 – continued

Formula	Stable Distortion	$\Delta H_{\text{stab}}^{\text{ABO}_3}$	ΔE_v^{O}	Ox. N.$_A$	Ox. N.$_B$
		[eV/atom]	[eV/O atom]		
$TlUO_3$	**cubic**	**-0.105**	**4.189**	**1**	**5**
$HoCrO_3$	orthorhombic	-0.033	4.191	3	3
$BaPuO_3$	rhombohedral	0.010	4.211	unknown	
$YbMoO_3$	**orthorhombic**	**0.007**	**4.239**	**2**	**4**
$SmVO_3$	orthorhombic	-0.064	4.244	2	4
$DyCrO_3$	orthorhombic	0.003	4.245	3	3
$YCrO_3$	orthorhombic	-0.020	4.254	3	3
$TmGaO_3$	**orthorhombic**	**0.019**	**4.279**	**3**	**3**
$PmVO_3$	**orthorhombic**	**-0.069**	**4.301**	**unknown**	
$TbCrO_3$	orthorhombic	-0.047	4.319	3	3
$NaReO_3$	**cubic**	**0.007**	**4.345**	**1**	**5**
$NaUO_3$	orthorhombic	0.003	4.356	1	5
$EuGeO_3$	**orthorhombic**	**-0.216**	**4.370**	**2**	**4**
$GdCrO_3$	orthorhombic	-0.052	4.371	3	3
$AgPaO_3$	**rhombohedral**	**-0.295**	**4.373**	**unknown**	
$NdVO_3$	orthorhombic	-0.077	4.379	3	3
$UAlO_3$	**cubic**	**-0.004**	**4.389**	**3**	**3**
$BaNpO_3$	orthorhombic	-0.088	4.442	unknown	
$PrVO_3$	orthorhombic	-0.083	4.464	3	3
UVO_3	**orthorhombic**	**0.021**	**4.467**	**3**	**3**
$YGaO_3$	orthorhombic	-0.090	4.492	3	3
$NpVO_3$	**orthorhombic**	**-0.032**	**4.514**	**unknown**	
$RbUO_3$	cubic	-0.050	4.519	1	5
$SmCrO_3$	orthorhombic	-0.064	4.544	3	3
$CeVO_3$	orthorhombic	-0.090	4.548	3	3
$PuAlO_3$	**orthorhombic**	**-0.071**	**4.577**	**unknown**	
$GdGaO_3$	orthorhombic	0.007	4.590	3	3
$PmCrO_3$	**orthorhombic**	**-0.067**	**4.600**	**unknown**	
$LaVO_3$	orthorhombic	-0.083	4.615	3	3
$NdCrO_3$	orthorhombic	-0.075	4.684	3	3
$NaMoO_3$	**cubic**	**-0.004**	**4.704**	**1**	**5**
$DyTiO_3$	orthorhombic	0.022	4.721	2	4
$SmGaO_3$	**orthorhombic**	**0.019**	**4.741**	**3**	**3**
$PrCrO_3$	orthorhombic	-0.079	4.772	3	3
$PmGaO_3$	**orthorhombic**	**-0.023**	**4.795**	**unknown**	
$TbTiO_3$	orthorhombic	0.015	4.803	2	4
$CeCrO_3$	orthorhombic	-0.081	4.838	3	3

Table 4.3 – continued

Formula	Stable Distortion	$\Delta H_{\mathrm{stab}}^{\mathrm{ABO_3}}$ [eV/atom]	ΔE_v^{O} [eV/O atom]	Ox. N.$_A$	Ox. N.$_B$
CeErO$_3$	**orthorhombic**	**0.023**	**4.848**	**3**	**3**
LaCrO$_3$	orthorhombic	-0.071	4.851	3	3
NdGaO$_3$	orthorhombic	-0.006	4.867	3	3
GdTiO$_3$	orthorhombic	0.023	4.871	2	4
NpCrO$_3$	**orthorhombic**	**-0.021**	**4.872**	**unknown**	
EuMoO$_3$	**orthorhombic**	**-0.119**	**4.931**	**2**	**4**
LaGaO$_3$	orthorhombic	0.015	4.940	3	3
PrGaO$_3$	orthorhombic	-0.056	4.942	3	3
CeTmO$_3$	orthorhombic	0.015	4.959	3	3
NdLuO$_3$	**orthorhombic**	**0.022**	**4.969**	**3**	**3**
AcVO$_3$	**orthorhombic**	**-0.106**	**4.974**	**unknown**	

4.4. Summary and Conclusions

Going beyond binary oxides for thermochemical water splitting applications opens a large composition space that is unreasonably big to be entirely explored experimentally. In this work, we used high-throughtput density functional theory to screen ABO$_3$ perovskites based on thermodynamic considerations. We did an exhaustive search of the all the possible ABO$_3$ combinations, without filtering for charge neutrality prior to performing the calculations which lead to the discovery of stable perovskites that have hard-to-predict oxidation states. We used two filters, compounds stability and oxygen vacancy formation energy, to isolate potential candidates for water splitting. The stability filter showed the importance of considering all competing phases present in the ternary A-B-O phase diagram to assess the stability of a compounds accurately. We found the majority of the stable perovskites to be orthorhombic with rare earth elements on the A-site and 3d-transition metals on the B-site. Plotting stable and unstable compounds in structural maps and computing their tolerance factor lead to the conclusion that purely geometrical argument are not sufficient to describe completely the formability of perovskites. Finally, we identified 139 perovskites that are predicted to be thermodynamically favorable for water splitting applications, some of those not reported in the literature. The high-throughput methodology presented in this paper shows the benefit of using first-principles calculations to efficiently screen an exhaustively large number of compounds at once. It provides a baseline for further studies involving more detailed exploration of a restricted number of those compounds.

CHAPTER 5

The Uniquely Large Entropy of Reduction of Ceria

Previous studies have shown that a large solid-state entropy of reduction increases the thermodynamic efficiency of metal oxides, such as ceria (CeO_2), for two-step thermochemical water splitting cycles (TWSC). The configurational entropy arising from oxygen off-stoichiometry in the oxide has been the focus of most previous work on the entropy of TWSC. Here we examine a different source of entropy, the onsite electronic configurational entropy ($\Delta S_{\mathrm{elec}}^{\mathrm{onsite}}$), arising from coupling between orbital and spin angular momenta ($L-S$) in lanthanide f-orbitals. We find that $\Delta S_{\mathrm{elec}}^{\mathrm{onsite}}$ is sizable in all lanthanides, and reaches a maximum value of $\approx 4.7\,k_{\mathrm{B}}$ per oxygen vacancy for the $\mathrm{Ce}^{+4}/\mathrm{Ce}^{+3}$ reduction reaction. Depending on the degree of non-stoichiometry in ceria, this value can even surpass the configurational entropy. The unique and large positive $\Delta S_{\mathrm{elec}}^{\mathrm{onsite}}$ in ceria contributes to its excellent water-splitting performance as well as its superior properties for other high-temperature catalytic redox reactions. Our calculations also show that TbO_2 – generally $\mathrm{Tb}^{+4}/\mathrm{Tb}^{+3}$ based materials – have a high electronic entropy and thus could also be potential candidates for solar thermochemical reactions.

5.1. Introduction

In the previous chapters, we mentioned ceria (CeO_2) as being the current best materials for water splitting.[3,11,23,150] In addition, several compounds from Table 4.3 contain cerium. Apart from water splitting, ceria is also used for various catalytic and energy applications such as three-way exhaust automotive catalysts,[151–155] solid-state fuel cells,[156–159] low-temperature water-gas shift reactions,[160] and several other industrial catalytic applications.[161–165] To a large extent, the performance of ceria in these processes depends strongly on its oxygen storage capacity and facile $\mathrm{Ce}^{+4}/\mathrm{Ce}^{+3}$ redox reaction. For water splitting, a critical step is the thermal reduction of ceria at around 2000 K:

$$\mathrm{MO}_x \rightarrow \mathrm{MO}_{x-\delta} + \frac{\delta}{2}\mathrm{O}_2 \tag{5.1}$$

where M is a metal, MO_x its corresponding metal oxide and δ is the oxygen off-stoichiometry. Ideally, for equation 5.1 to be thermodynamically favorable, its Gibbs free energy has to be negative:

$$\Delta G_{\mathrm{red}} = \Delta H_{\mathrm{red}} - T_{\mathrm{red}}\Delta S_{\mathrm{red}} < 0 \tag{5.2}$$

where ΔH_{red} is the enthalpy of reduction, which we discussed in chapter 4, T_{red} is the reduction temperature (typically around 2000 K) and ΔS_{red} is the entropy of reduction. Meredig and Wolverton[125] showed that a key thermodynamic quantity for increase efficiency is a large ΔS_{red}. This entropy of reduction for a thermochemical water splitting process is conventionally defined as:

$$\Delta S_{\mathrm{red}} = \frac{1}{2}S_{\mathrm{O}_2} + \Delta S_{\mathrm{conf}} + \Delta S_{\mathrm{vib}} \tag{5.3}$$

where S_{O_2} is the oxygen gas phase entropy, ΔS_{conf} is the ionic and electronic configurational entropy and ΔS_{vib} is the vibrational entropy. The oxygen gas phase is independent from the metal oxide used in equation 5.1 and is approximately $15\,k_{\mathrm{B}}$ per oxygen atom.[166,167] The two other terms of equation 5.3 are the materials dependent quantities and are referred as the solid-state entropy of reduction, $\Delta S_{\mathrm{red}}^{\mathrm{solid}}$.

Several studies tackled the solid-state entropy. Experimentally, Bevan *et al.*[168] and Panlener *et al.*[166] showed that, by using an ideal solution model, $\Delta S_{\mathrm{red}}^{\mathrm{solid}}$ is logarithmically dependent on the oxygen off-stoichiometry, δ. Grieshammer *et al.*[169] calculated the vibrational entropy to be equal to $2.5\,k_\mathrm{B}$. Gopal *et al.*[170] did Monte Carlo (MC) simulations based on DFT-derived cluster expansion Hamiltonian to calculate the configurational and vibrational entropy of reduction of ceria for various δ. They found that the configurational entropy is much smaller than the ideal solution model. Their values of entropy of reduction agree with experiment for large δ but have a gap of about $4.5\,k_\mathrm{B}$ for small δ ($0.01 < \delta < 0.12$).

To explain this discrepancy, we will look at an additional source of entropy, hereafter denoted by $\Delta S_\mathrm{elec}^\mathrm{onsite}$, which arises from distributing electrons over a large number of multiplet states. This onsite electronic entropy is particularly large for lanthanides with partially filled f-shells where extremely localized f-orbitals give rise to different possible configurations associated with the occupations of the same atomic orbitals. In addition, we calculate the onsite electronic entropy for different lanthanides cations (Pr, Nd, Eu and Tb) and show that the onsite electronic entropy is the largest for the $\mathrm{Ce}^{+4} \rightarrow \mathrm{Ce}^{+3}$ reduction reaction, explaining the unique properties of ceria.

5.2. Results and Discussion

5.2.1. L-S Coupling and Crystal Field

The onsite electronic entropy arises from thermal excitations among orbitals created by orbital angular momentum (L) and spin angular momentum (S) coupling ($L{-}S$ coupling). For f-orbitals of lanthanides, we use the Russel-Saunders ($L - S$) coupling scheme[171] to describe the electronic configuration instead of the number of valence electrons ($4f^n$) notation. In this scheme, coupling of orbital and spin angular momentum results in ${}^{2S+1}L_J$

term symbol in which $2S+1$ is the spin-multiplicity, L is the total orbital quantum number and J is the total angular momentum, ranging from $|L+S|$ to $|L-S|$ by steps of one. The degeneracy of each J-multiplet is $(2J+1)$ and the total number of microstates (m) for a given term symbol ${}^{2S+1}L$ is $(2S+1)(2L+1)$.

When a cation is placed in a crystal where it is surrounded by anions, static electric field breaks the degeneracy of electron orbitals (crystal field (CF)).[172] In our system, CF further splits each degenerate J-state to several subsets and breaks the spherical symmetry of the f-shell charge distribution. The crystal field parameters are dependent on the local symmetry of the ionic environment. Hence, we used a fully *Ab initio* method, opposing crystal potential (OCP),[173] to calculate the CF parameters of Ce^{+3} in the host fluorite CeO_2 structure. Figure 5.1 shows the $4f^1$ (Russel-Saunders notation: ${}^2F$) energy level splitting scheme of Ce^{+3} with SOC and calculated crystal field. Without CF, the f^1 states split into ${}^2F_{5/2}$ and ${}^2F_{7/2}$ separated by approximately 0.28 eV.[174] The CF interaction further splits the 6-fold degenerate ${}^2F_{5/2}$ ground state into a four-fold degenerate Γ_8 and two-fold degenerate Γ_7 subsets, separated by 0.12 eV. Crystal field, which was calculated by OCP method for Ce^{+3}, splits the eight-fold degenerate ${}^2F_{7/2}$ state into states with energies 0.25, 0.32 and 0.46 eV.

5.2.2. Onsite Electronic Entropy

Once the energy levels and degeneracy of each microstates are known, we can calculate the onsite electronic entropy of the system as follow:

$$S_{\text{elec}}^{\text{onsite}} = -k_{\text{B}} \sum_{i}^{m} g_i \, p_i \, \ln \, p_i \tag{5.4}$$

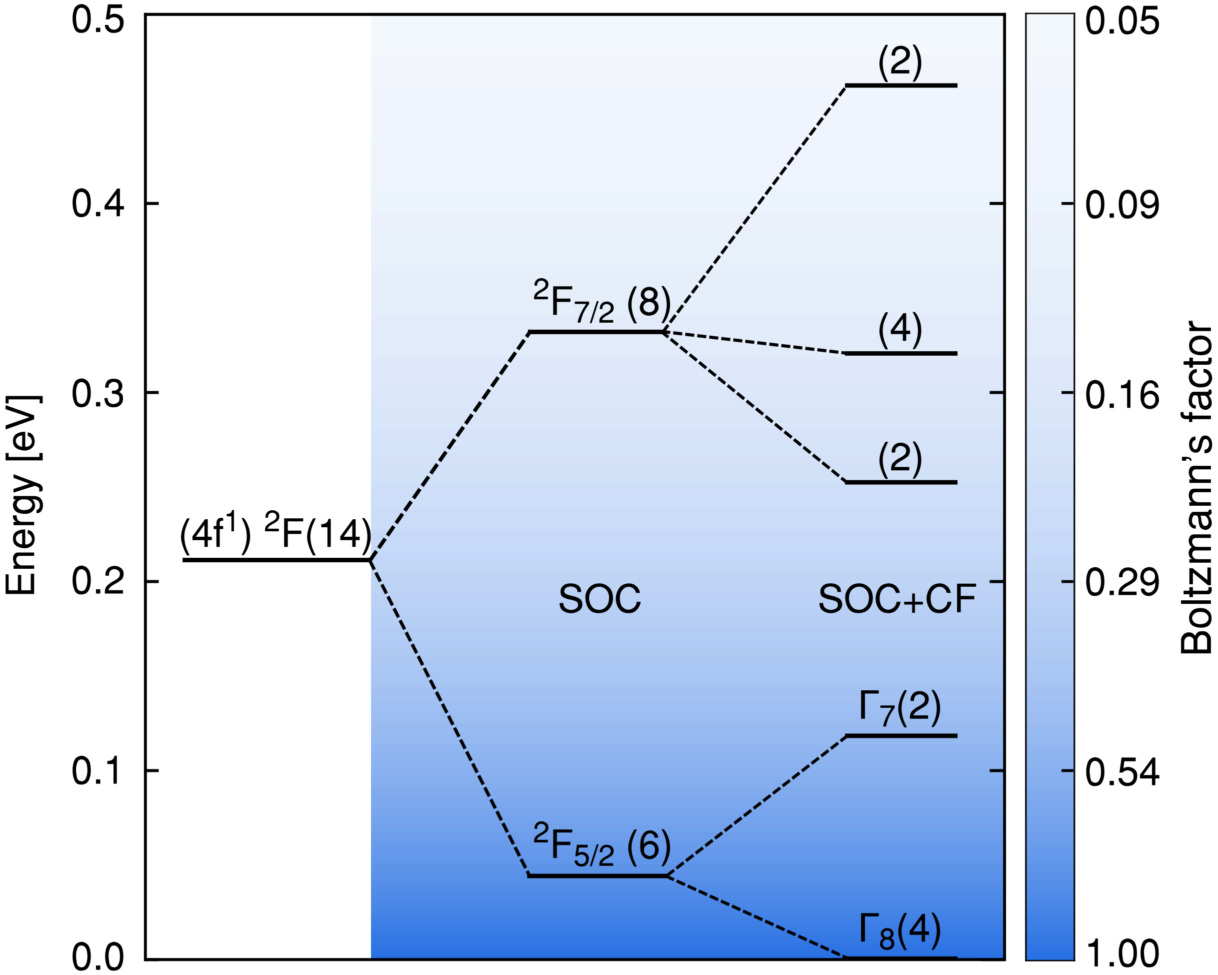


Figure 5.1. Energy levels of the $4f^1$ orbital of Ce^{+3}. Ce^{+3} splits initially by SOC and subsequently by cubic CF of the the fluorite structure. The spin-orbit splitting between $J = 5/2$ and $J = 7/2$ is about 0.28 eV.[174,175] The color gradient indicates the probability distribution at 1900 K, given by $\exp(-E_i/\mathrm{k_B}T)$, and numbers in parentheses stand for the degeneracy of the electronic states. The first predicted Γ_8 to Γ_7 excitation for CeO$_2$ is 0.12 eV. Predictions for the higher CF levels of $J = 7/2$ are 0.25, 0.32, 0.46 respectively.

where k_B is the Boltzmann factor ($8.617 * 10^{-5}$ eV/K), m is the number of microstates, g_i the degeneracy of the microstate m_i and p_i is the probability of thermal excitation to the state with energy E_i given by:

$$p_i = \frac{\exp(-E_i/k_\mathrm{B}T)}{Z} \tag{5.5}$$

where T is the temperature and Z is the partition function defined as:

$$Z = \sum_{i}^{m} g_i \exp(-E_i/k_\mathrm{B}T) \tag{5.6}$$

Equations 5.4-5.6 show that the onsite electronic entropy depends mainly on the number of microstates (m) and the probability of occupying them. This probability is dependent on the temperature and the size of the multiplet splitting between the energy levels: stronger SOC means higher energies microstates that are less probable to be occupied at lower temperatures due to limited thermal excitations. However, for Ce^{+3} at temperatures relevant for water splitting (T $\approx$ 1900 K) a large fraction of microstates are accessible making the $S_\mathrm{elec}^\mathrm{onsite}$ close to its ideal limit of $k_\mathrm{B}\ln(m)$.

Table 5.1 contains the onsite electronic entropy for the 5 elements (Ce, Pr, Nd, Eu and Tb) that were considered in this study. Myers *et al.*[176] extracted the electronic entropy contribution of lanthanide ions (Ln^{+3}) in lanthanide trihalides from absolute entropy data. Our calculated electronic entropies per ion at $\approx$ 300 K in units of k_B compared to Myers *et al.*[176] data (value inside parentheses) are the following: Ce^{+3}, 1.79 (1.77); Pr^{+3}, 2.19 (2.18), Nb^{+3}, 2.30 (2.27); Eu^{+3}, 1.13 (1.10); Tb^{+3}, 2.56 (2.54). The calculated $S_\mathrm{elec}^\mathrm{onsite}$ based on $L - S$ coupling shows excellent agreement with previously reported data.

Table 5.1. Calculated onsite electronic entropy per oxygen vacancy, $S_{\text{elec}}^{\text{onsite}}$, of selected lanthanide ions before and after reduction at 1900 K. Once the f-orbitals are occupied, the system gains a large electronic entropy which weakly depends on its occupation number. Therefore, the largest $\Delta S_{\text{elec}}^{\text{onsite}}$ per oxygen vacancy is associated with the f^0 to f^1 transition, where fully oxidized state has zero entropic contribution. Entropy units are in k_{B}.

Element	f^n	Term	Deg.	$S_{\text{elec}}^{\text{onsite}}$	$\Delta S_{\text{elec}}^{\text{onsite}}$
Ce^{+4}	f^0	^1S	1	0.0	4.68
Ce^{+3}	f^1	^2F	14	4.68 $(4.53)^{\text{CF}}$	
Pr^{+4}	f^1	^2F	14	4.38 $(4.22)^{\text{CF}}$	1.40
Pr^{+3}	f^2	^3H	33	5.78	
Nd^{+3}	f^3	^4I	52	6.28	0.77
Nd^{+2}	f^4	^5I	65	7.05	
Eu^{+3}	f^6	^7F	49	6.59	-2.43
Eu^{+2}	f^7	^8S	8	4.16	
Tb^{+4}	f^7	^8S	8	4.16	2.30
Tb^{+3}	f^8	^7F	49	6.46	

For thermochemical water splitting applications, the absolute electronic entropy does not matter, only the entropy difference before (f^{n}) and after ($f^{\text{n+1}}$) reduction is relevant:

$$\Delta S_{\text{elec}}^{\text{onsite}} = 2\big(S_{\text{elec}}^{\text{onsite(n}-1)} - S_{\text{elec}}^{\text{onsite(n)}}\big) \tag{5.7}$$

where the factor two is due to the fact that two Ce^{+4} ions are reduced per oxygen vacancy. Table 5.1 shows the onsite electronic entropy of reduction for all the elements considered. The largest $\Delta S_{\text{elec}}^{\text{onsite}}$ is found in $\text{Ce}^{+4} \rightarrow \text{Ce}^{+3}$ which undergoes an $f^0 \rightarrow f^1$ redox reaction. Indeed, having the oxidized state f^0 (^1S) with zero onsite electronic entropy is a unique feature of ceria, resulting in a large $\Delta S_{\text{elec}}^{\text{onsite}}$ of 4.68 k_{B} per oxygen vacancy, which is a maximum for the reduction of any rare-earth cation. We assert that this unique entropic characteristic of the Ce^{+4}/Ce^{+3} redox reaction helps facilitate the TWSC properties of CeO_2. The second largest value of $\Delta S_{\text{elec}}^{\text{onsite}}$ is found in terbium ($Tb^{+4} \rightarrow Tb^{+3}$) with $2.30\,k_{\text{B}}$ per oxygen vacancy at 1900 K. This source of entropy could make Tb^{+4} based materials promising candidates for TWSC applications, as Tb, like Ce, is stable in two valence states (Tb^{+4}/Tb^{+3}). This prediction agrees with a recent thermodynamic study that also suggested TbO_2 as a potential candidate for TWSC applications.[177]

For cerium and praseodymium, we calculated $S_{\text{elec}}^{\text{onsite}}$ with and without crystal field splitting and showed that it has a negligible contributions at temperatures relevant for water splitting ($\approx 3\%$ at 1900 K, see Table 5.1 and Figure 5.2). In addition, Walsh *et al.*[178] showed that crystal field splitting significantly decrease with temperature and lattice thermal expansion. As a result, the CF parameters were not calculated for the remaining elements.

5.2.3. Other Sources of Entropy

In this section, we compare $\Delta S_{\text{elec}}^{\text{onsite}}$ with the other sources of entropy. With the addition of onsite electronic entropy, equation 5.3 becomes:

$$\Delta S_{\text{red}} = \frac{1}{2} S_{O_2} + \Delta S_{\text{conf}} + \Delta S_{\text{vib}} + \Delta S_{\text{elec}}^{\text{onsite}} \tag{5.8}$$

For simplicity we consider a fixed composition of $\delta = 0.03$ roughly corresponding to one oxygen vacancy in a 96-atom supercell. For this composition, we were able to find several reported experimental and theoretical data points (Table 5.2).

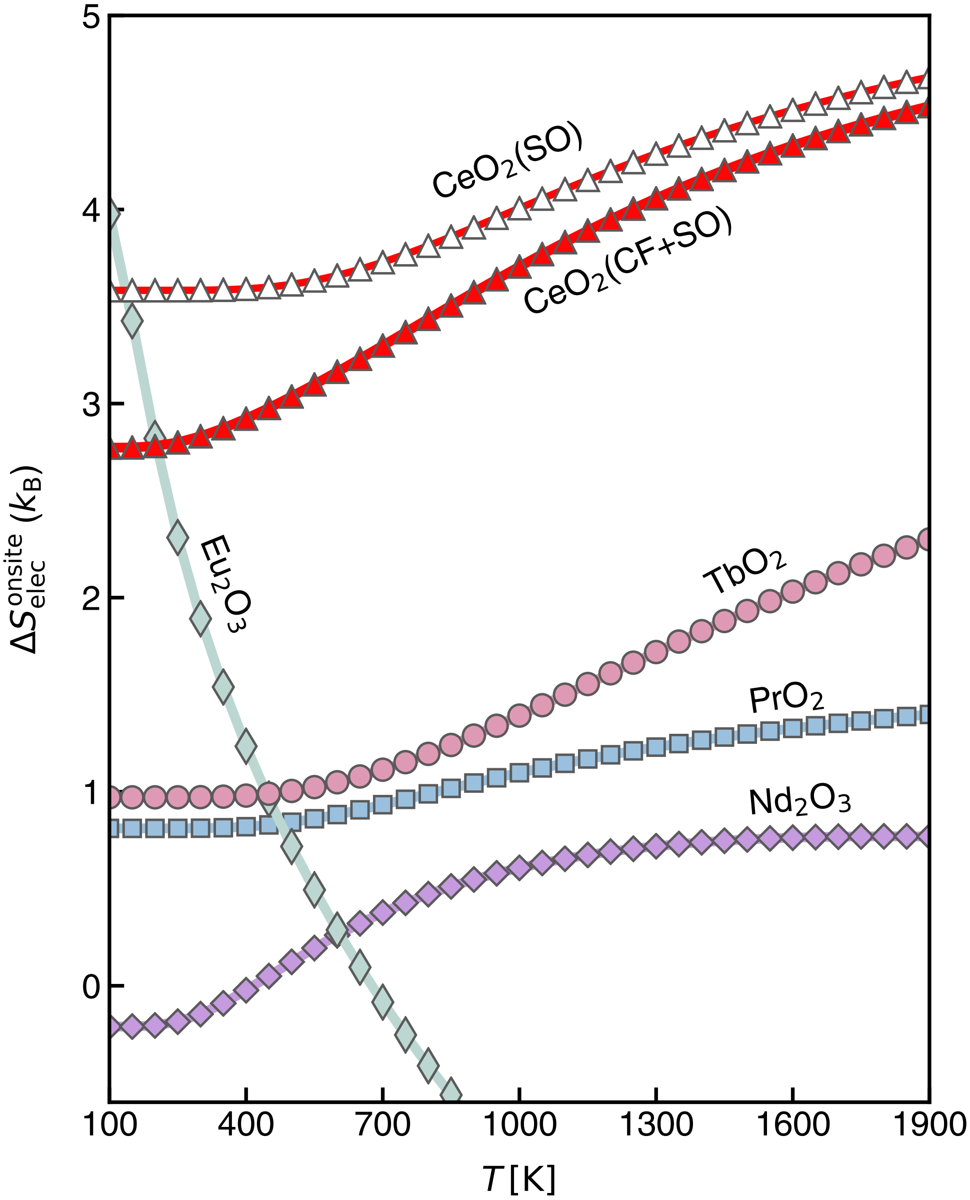


Figure 5.2. Calculated $\Delta S_{\mathrm{elec}}^{\mathrm{onsite}}$ for lanthanides ions in function of temperature. Predicted electronic entropy of reduction per oxygen vacancy for the lanthanide oxides studied in this work. At high temperature, reduction of CeO_2 has the highest $\Delta S_{\mathrm{elec}}^{\mathrm{onsite}}$ followed by reduction of TbO_2 (see Table 5.1).

Table 5.2. Contribution of different entropic terms for $\delta = 0.03$ and a temperature of 1500 K. The values of ΔS_{conf} are obtained from an ideal solution model and MC simulations:[170] the MC calculated S^{conf} already includes vibrational entropy.[170] Experimental value is taken from Panlener *et al.*[166]

Method	$\frac{1}{2}S^0_{O_2}$	ΔS^{vib}	ΔS^{conf}	$\Delta S^{\mathrm{onsite}}_{\mathrm{elec}}$	ΔS^{tot}	$\Delta S^{\mathrm{tot}}_{\mathrm{exp}}$
Ideal	15.2	2.5	10.4	—	28.1	26.1
MC	15.2	5.9		—	21.1	
MC+$\Delta S^{\mathrm{onsite}}_{\mathrm{elec}}$	15.2	5.9		4.26†	25.4	

† This value is calculated for T=1500 K

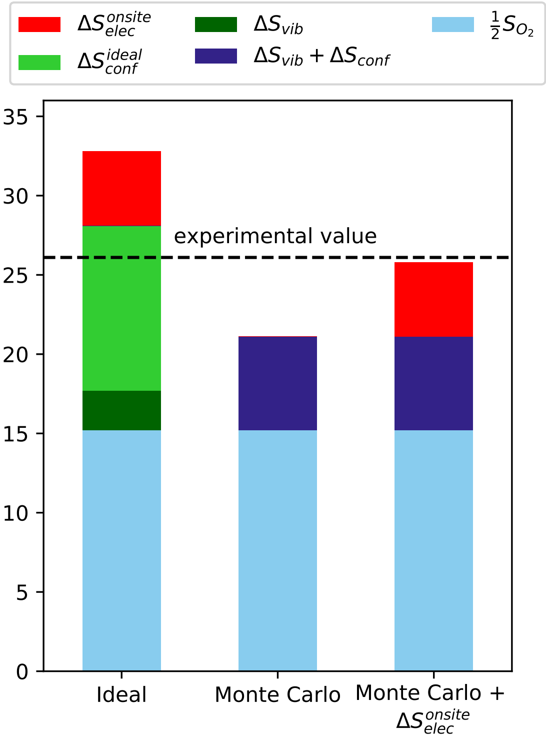


Figure 5.3. Contribution of different entropic terms for $\delta = 0.03$ and 1500 K. All the numbers are taken from Table 5.2.

At this composition the calculated ΔS_{vib} is approximately 2.5 k_{B}.[169] The ΔS_{conf} of $\mathrm{CeO}_{2-\delta}$, assuming ideal mixing entropy is calculated by $\Delta S_{\mathrm{c}} = -nk_{\mathrm{B}}\ln(\delta)$ (where n depends on the defect structure, here $n = 3$) and is equal to 10.4 k_{B}.[166,170] However, we note that a system with extensive ordering of oxygen vacancies,[147] such as ceria, will have short range order and hence the actual configurational entropy is non-ideal and smaller than in the ideal solution model. For instance, the non-ideal $\Delta S_{\mathrm{conf}} + \Delta S_{\mathrm{vib}}$, calculated by Monte Carlo simulation based on a cluster expansion Hamiltonian, is about $5.9\,k_{\mathrm{B}}$,[170] almost half of the ideal ΔS_{conf}. Our calculations show that the neglected electronic entropy (ΔS_{elec}) is more than $4.7\,k_{\mathrm{B}}$, which is comparable to these other widely considered sources of entropy and can explain the $\approx 5\,k_{\mathrm{B}}$ gap between the calculation and experiment. We note that as long as oxygen vacancy is compensated by two polarons (i.e. $[\mathrm{Ce}'_{\mathrm{Ce}}] = 2[\mathrm{V}_{\mathrm{O}}^{\bullet\bullet}]$), $\Delta S_{\mathrm{elec}}^{\mathrm{onsite}}$ is not a function of the off-stoichiometry δ. Being independent of the off-stoichiometry implies that at large δ the contribution from the electronic entropy surpasses that of the configurational entropy (which decreases with δ) and becomes the major entropic contribution. Using the calculated ΔS_{conf} in ref. 170, we estimate that this crossover occurs at ($\delta \approx 0.05$).

Our results show that the electronic contribution to the entropy of reduction explains the gap between the results of the currently most detailed theoretical calculations of ref. 170 and the experimental data of Panlener *et al.*[166] for small δ (see Figure 5.3). At larger δ, adding a constant onsite electronic entropy to the vibrational and configurational entropies from ref. 170 overestimates the experimental data (see Figure 5.3). There could be several reasons for this apparent discrepancy. For instance, at higher δ, most of the polarons become bound to oxygen vacancies forming singly charged $\mathrm{V}_{\mathrm{O}}^{-2}$– Ce^{+3} or neutral $\mathrm{V}_{\mathrm{O}}^{-2}$–$2\mathrm{Ce}^{+3}$ complexes.[179] The proximity of Ce^{+3} to an oxygen vacancy could slightly modify the electronic structure and hence reduce the electronic entropy associated with Ce^{+3}, but as already discussed the overall effect of oxygen vacancy on the energy levels[180] and electronic entropy is expected to be small. Furthermore, the experimental measurements of Panlener *et al.*[166] found that the enthalpy of reduction is composition dependent even at very small δ. However, this finding has been challenged due to the large experimental uncertainty.[166,168] Since the entropy is obtained from $T\Delta S = \Delta H - \Delta G$, the entropy values of Panlener *et al.* may be contaminated by contributions from the composition dependent contribution to ΔH. Indeed, the results of ref. 170 suggest that the entropy stays approximately constant for δ between 0.05 and 0.15, while the data of Panlener *et al.*[166] shows a pronounced decrease in this range.

Measurements of the Seebeck coefficient provide another means of estimating the electronic entropy contribution in the dilute limit where all polarons are unbound.[151,181] Unfortunately, the experimental data here are also contradictory. The data of Tuller and Nowick[181] suggests that for small δ the spin degeneracy factor is one, which contradicts the Kramers theorem requiring that the ground state must be at least doubly degenerate. However, a later study by the same authors[151] concluded that the agreement between the polaron model with spin degeneracy one and the experimental data for the Seebeck coefficient was poor, especially at low δ where impurities were thought to play an important role. On the theory side, the vibrational entropy of an isolated Ce^{+3} polaron has not been established accurately. Grieshammer *et al.*[169] have calculated a very large value of about $7\,k_\mathrm{B}$ for the entropy of polaron formation at zero pressure, but the largest contribution to this value is due to a volume contribution from the CeO_2 host, which was treated in an approximate fashion. Such a large positive entropy is inconsistent with the available data on the Seebeck coefficients in the dilute limit.[151,181] Hence, thermoelectric measurements on pure, well-equilibrated samples of CeO_2 and more accurate calculations of the vibrational entropy associated with free polarons are highly desirable.

5.3. Conclusions

We calculated electronic entropies of different lanthanides in the presence of SOC and CF. We calculated CF splittings for Ce^{+3} and Pr^{+4} and found that at temperatures above 1000 K, CF interactions affect the S_{elec} by less than 3%. The results show that, in ceria, the magnitude of the entropy of reduction due to the commonly neglected onsite electronic entropy (ΔS_{elec}) reaches a maximum of $4.68\,k_{\mathrm{B}}$ per oxygen vacancy, which is twice as large as the vibrational entropy contribution and can be larger than the configurational entropy. This surprisingly large entropy is the result of the very unique electronic structure of cerium in ceria where redox reactions change its electronic state from f^0 to f^1. These entropic properties, together with the excellent chemical stability and tolerance for large non-stoichiometry, put ceria in a unique position for two-step solar thermochemical CO_2/H_2O splitting cycles. In addition, we find that Tb(IV) based materials have the next highest electronic entropy, for $Tb^{+4} \rightarrow Tb^{+3}$ redox reactions. We therefore propose compounds containing Tb^{+4} should be experimentally investigated as promising candidates for TWSC applications.

CHAPTER 6

Optimizing Machine Learning Methods for Faster Materials Discovery

In the recent years, machine learning has been used in materials science to predict materials properties and to accelerate the discovery of new stable compounds. Here, we take a critical look at methods of discovering new crystalline compounds to find ways to improve their performance. Specifically, we use an exhaustive dataset of ABO_3 compounds to test different training set types, algorithms and iterative schemes in order to improve each step of the materials discovery process. We show that building a training set from data coming exclusively from literature is not always necessary and can even be detrimental to the discovery rate of new compounds. In addition, we show that an iterative search approach, where unknown compounds are continuously calculated and included in the training set, lead to faster short-term discovery of compounds. Finally, we give a roadmap to perform machine learning for materials discovery in an efficient way.

6.1. Introduction

With the progress in computational power of the recent decades, it is now possible to calculate materials properties from first-principles in a high-throughput fashion. The databases resulting from such methods contain hundreds of thousands of materials properties, such as thermodynamic stability, relaxed geometries and band gaps.[59,62,129,182] Density functional theory (DFT), the current workhorse of such databases, remains expensive and is thus limited to structures with small number of atoms (typically < 50 for high-throughput calculations) or constrained chemical spaces. Additionally, in materials science, the first screening criteria is often stability, i.e. the likelihood of synthesizing a compound experimentally.[117] Unfortunately, stable compounds often represent a small fraction of the total number of compounds that are being calculated.[62,183–185] As a result, there is a benefit to reducing the number of calculations performed on unstable compounds which would lead to acceleration in materials discovery by allowing to explore bigger cells over the entire periodic table of the elements.

The sheer amount of data available, most often publicly, represents unique opportunities for machine learning and data mining for materials science.[75,85–87,186,187] Previous studies used such data to predict a variety of properties such as crystal structures,[64,81,188,189] melting temperatures[80,190] and mechanical properties of materials.[82,191,192] Machine learning has also frequently been used to discover new crystalline materials.[83,184,193] A common way to perform materials discovery with machine learning is to first build a training set by sampling experimentally observed compounds as those data are readily available in the literature.[81,194,195] As a result, the training sets are often biased towards positives, i.e. the ratio stable/non-stable in the training set is much higher than the one of the entire chemical space. This can potentially hurt the prediction capabilities as a biased training set will have the tendency to predict a higher number of false positive compounds. In addition, positive examples of stability are materials being stable in the desired structure and negative examples are materials stable in a different structure. As a result, there are very few examples of unstable materials, i.e. materials that decomposes in more than one phase.

One way to verify the predictions of machine learning based on literature training sets is to try to synthesize the predicted compounds, which is expensive when many materials are predicted to be stable. An alternative way of assessing the machine learning predictions is to validate the stability by performing ab-initio calculations, such as density functional theory. However, by doing so, there is a discrepancy between the labels of the training set (stable/non-stable in the literature) and the verification of the predictions (stable/non-stable as predicted by DFT). The connection between those two quantities is not always apparent.

In this work, we explore how to further improve each step each in the materials discovery process by performing machine learning on a complete dataset of ABO_3 compounds computed by DFT. Having a full dataset of 5,329 compounds, will allow us to compare the effect of different algorithms, training sets and machine learning approaches. We show that using machine learning can increase the discovery rate of new stable compounds by a factor 10. Furthermore, the machine learning results indicate that having data from the literature as a training set is not always necessary. Indeed, efficient machine learning can be performed with training set containing a random selection of compounds within the chemical space of interest. Finally, we try a different version of active learning[87,196–198] aimed at discovering new stable compounds as fast as possible while spending a minimum of computer time calculating unstable compositions. The iterative search method showed in this work uses a greedy approach to compute compounds with a high prediction score first.

6.2. Methodology

6.2.1. The Datasets

In a previous work, we calculated the thermodynamic stability of 5,329 ABO_3 compounds with DFT[68] by substituting 73 metals and semi-metals on both the A and B sites ($73^2 =$ 5,329). Out of those, 383 ($\approx$ 7%) were perovskites. This considerable CPU time expenditure gives us the unique opportunity of having an exhaustive and consistent dataset that can be used for machine learning purposes. In the present work, we use a subset of 65 elements that have been reported in the literature in a ABO_3 structure,[110,138–140] resulting in $65^2 =$ 4,225 different compositions. Out of those, 305 ($\approx$ 7%) are predicted by DFT to be stable perovskites. The breakdown of elements used in this study is summarized in Figure 6.1. By already knowing whether a compound is predicted by DFT to be stable or not, we can split this dataset into a training and testing set without restrictions, i.e. without being constrained by having only literature data. Here, we refer to the training set as the subset of compounds that will be used to train the machine learning model and the testing set as the remaining compounds, i.e. compounds that the machine learning model has not seen. To mimic what is typically done in other studies[81,184,194,195] and to study the effects of training set selection on machine learning performance, we will perform two types of splitting that are represented in Figure 6.2.

Literature Training Set. One approach to starting a machine learning search for new compounds is to gather all examples of experimentally observed compounds with the same stoichiometry, and train a model to predict whether the desired structure is formed.[81,194,199] For this work, the training is composed of 343 experimentally known

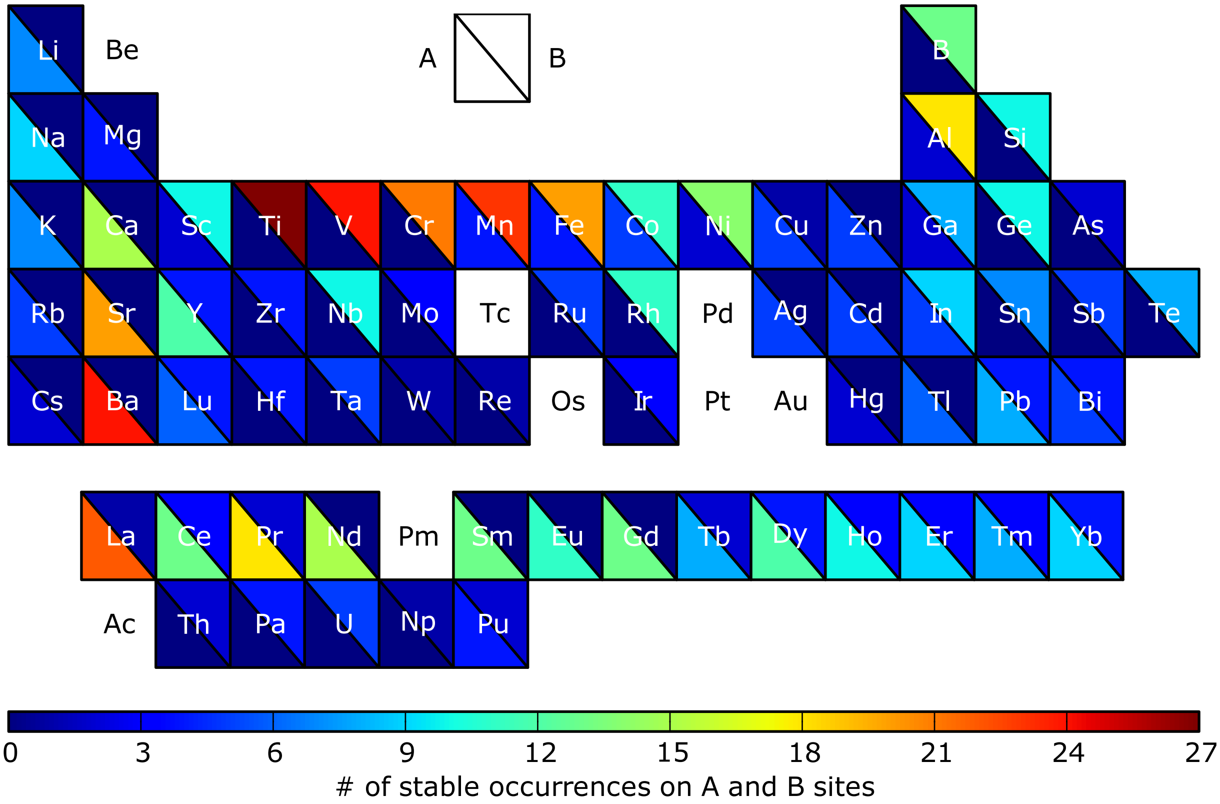


Figure 6.1. Elements found in the literature in ABO_3 compounds. Elements are color coded based on their occurrence on the A- and/or B-site. White elements with black symbols were not included in this study.

ABO_3 compounds found in four review papers.[110,138–140] To emulate the approaches found in literature, the compounds are then split according to their DFT stability. In order to be considered as a stable perovskite, an ABO_3 compound has to be on the convex hull in the perovskite structure, i.e. has a lower energy than any other ABO_3 non-perovskite compounds and any linear combinations of other phases in the A-B-O phase space.[62,68,132] For this splitting, we have 173 perovskites and 170 non-perovskites in the training set (see Figure 6.2 (a)).

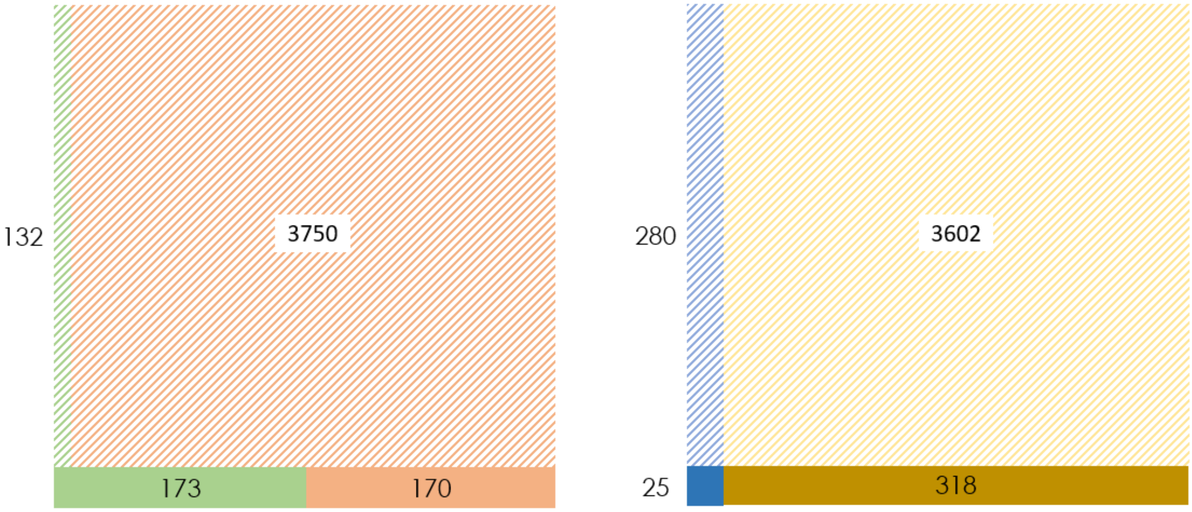


Figure 6.2. Dataset representation for both types of training set. Each large square represents $65 * 65 = 4225$ compositions. The bottom solid rectangles are the training sets, the top striped rectangles are the testing sets. (a) Splitting according to compounds found in the literature, the green and brown colors represent compounds predicted to be perovskite and non-perovskite, respectively. (b) Random selection of training set, the blue and yellow represent compounds predicted to be perovskite and non-perovskite, respectively.

Random Selection of the Training Set. Alternatively, we can compose the training set by randomly selecting 343 compounds (the same number as the literature training set) out of the 4,225 compositions, similar to the approach used by Faber *et al.*[184] This method can be used if the structure of interest is largely absent from the literature. Coupled with ab-initio tools such as DFT, it offers the advantage of not requiring any previous data to perform materials discovery. In addition, selecting the training set randomly guarantees that the ratio perovskite/non-perovskite is identical to the overall dataset (see Figure 6.2 (b)).

6.2.2. Feature Set and Target Values

In this work, we will use 8 attributes as inputs into our machine learning models. As all compositions have 3 oxygen atoms, all attributes are derived from the two cations present in the structure. Those attributes are: 1) Charge state of the ABO_3 compound. This is a binary attribute based on whether a compound is charged balanced or not. The oxidation state of the two cations in the structure was calculated using a bond valence method.[104] The oxidation state of oxygen was fixed to -2 for all compounds. 2-5) The column and row of the element of the periodic table for the A and B atom. 6-7) the atomic radii of the A and B atoms. Here, we will investigate the effect of two type of atomic radii: covalent and ionic radii. Covalent radii measure the size of an atom when forming a covalent bond[200] and are thus independent of the oxidation of coordination number of the element. In contrast, ionic radii are a measure of the atoms ion and are thus dependent on the oxidation state and coordination number of the ion.[106,107] 8) Tolerance factor[41] defined by equation 2.5. Tolerance factor is often used in the literature when describing the formability of perovskites.[39] The general consensus based on empirical observations of stable ABO_3 compounds is that $0.8 < t < 1.1$ for perovskites. We note that the octahedral factor, another metrics often encountered, defined as:

$$\mathcal{O} = \frac{r_B}{r_O} \tag{6.1}$$

Is already included in our feature set 6-7) as the oxygen radius is a constant.

The target value for this study is the DFT 0 K, 0 bar stability of the compound. It is a binary quantity that represents whether the compound is predicted to be stable by DFT or not. A phase is considered stable when its energy is lower than any other structures at that composition or any other linear combinations of structures. Grand canonical linear programing method (GCLP) is used to calculate the stability of every phase.[62,97]

6.2.3. Machine Learning Algorithm

Throughout this work, we will use different algorithms from scikit-learn,[79] the machine learning python package, to perform supervised learning. Specifically, we tested two different classes of learning algorithms.

Decision Trees Ensemble Methods. In order to improve the predictive accuracy of the model, ensemble methods use a combination of weak predictors to build the final model. In this work, we will use a random forest (RF)[201] and gradient boosting (GB)[202] decision trees classifiers which both use decision trees as weak classifiers. RF performs an average over decision trees build on a sub-sample of the dataset. While the sub-sample dataset sizes are always the same, they are created from different random bootstrap samples of the original training data. Gradient Boosting uses decisions trees that are built recursively to minimize the error between the residuals, i.e. the difference between the target function and the prediction. Both ensemble methods are probabilistic. As a result, we train a model and predict on our dataset 100 times to have statistically meaningful results.

Non-linear support vector machines (SVM). Non-linear SVM (or kernel SVM, k-SVM) uses kernel trick to implicitly map inputs to higher-dimension space in order to find a hyperplane that splits the data into two categories (here perovskite and non-perovskite). SVM are deterministic algorithms meaning that we train a model and predict with it only once. In this work, we use the C-vector classification[203] with radial basis function as kernel.

6.2.4. Performance Metrics

The performance of the different algorithms will be compared in terms of confusion matrix (Table 6.1). The accuracy of the algorithm assesses how many machine learning predictions agrees with DFT over the total number of predictions:

$$\text{accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FN} + \mathrm{FP}} \tag{6.2}$$

where TP, TN, FN and FP correspond for true positive, true negative, false negative and false positive, respectively. As a result of the sparse number of perovskites in the dataset, we note that guessing non-perovskite for every compound would yield to an accuracy of (4225-305)/4225 $\approx$ 93%. A metric that is more important for accelerating materials discovery is success rate of the predictions (positive predictive value), i.e. how many compounds are predicted to perovskite by DFT divided by the total amount of DFT calculations performed:

$$\text{success rate} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{6.3}$$

By doing DFT without machine learning on all the possible compositions, the success rate, referred as the random guessing rate, is 305/4225 $\approx$ 7%. The hope is that machine learning can increase this percentage.

Table 6.1. Confusion matrix. TP, TN, FN and FP stand for true positive, true negative, false negative and false positive, respectively.

		Machine learning predicts	
		perovskite	non-perovskite
DFT predicts	perovskite	TP	FN
	non-perovskite	FP	TN

6.2.5. Iterative Search Approach

Typically, machine learning is performed by training models on a subset of known data points (referred as the training set) and used to predict the properties of another set of compounds, usually larger, for which the target property is unknown (unknown set). With this “single iteration” approach, the quality of the predictions is often measured in function of the n-fold cross-validation score (where n is typically 10), i.e. partitioning the training set in n part, training a model on n-1 partition, using the model to predict the data points that were excluded from the training set and repeating this process n times. Even if cross-validation offers an insight on how the model should perform on the unknown dataset, it offers little validation on the predictions.

The single iteration approach only involves training a model and using it to predict compounds once. However, in the case of materials discovery, the search space might be too big to be calculated at once. Furthermore, the training set size, if taken from literature data, might be too small to offer accurate predictions. For those reasons, we test a form of iterative search where DFT calculations of unknown compounds are done successively and added to the training set continuously. The choice of which compounds to compute first is based on the prediction confidence made by the machine learning algorithm (compounds

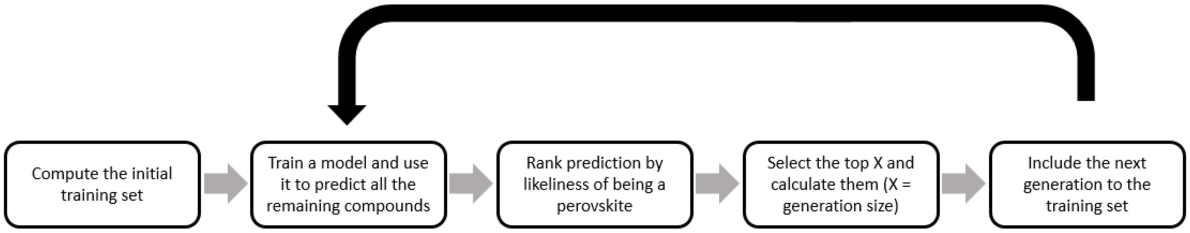

Compute the initial training set
Train a model and use it to predict all the remaining compounds
Rank prediction by likeliness of being a perovskite
Select the top X and calculate them (X = generation size)
Include the next generation to the training set

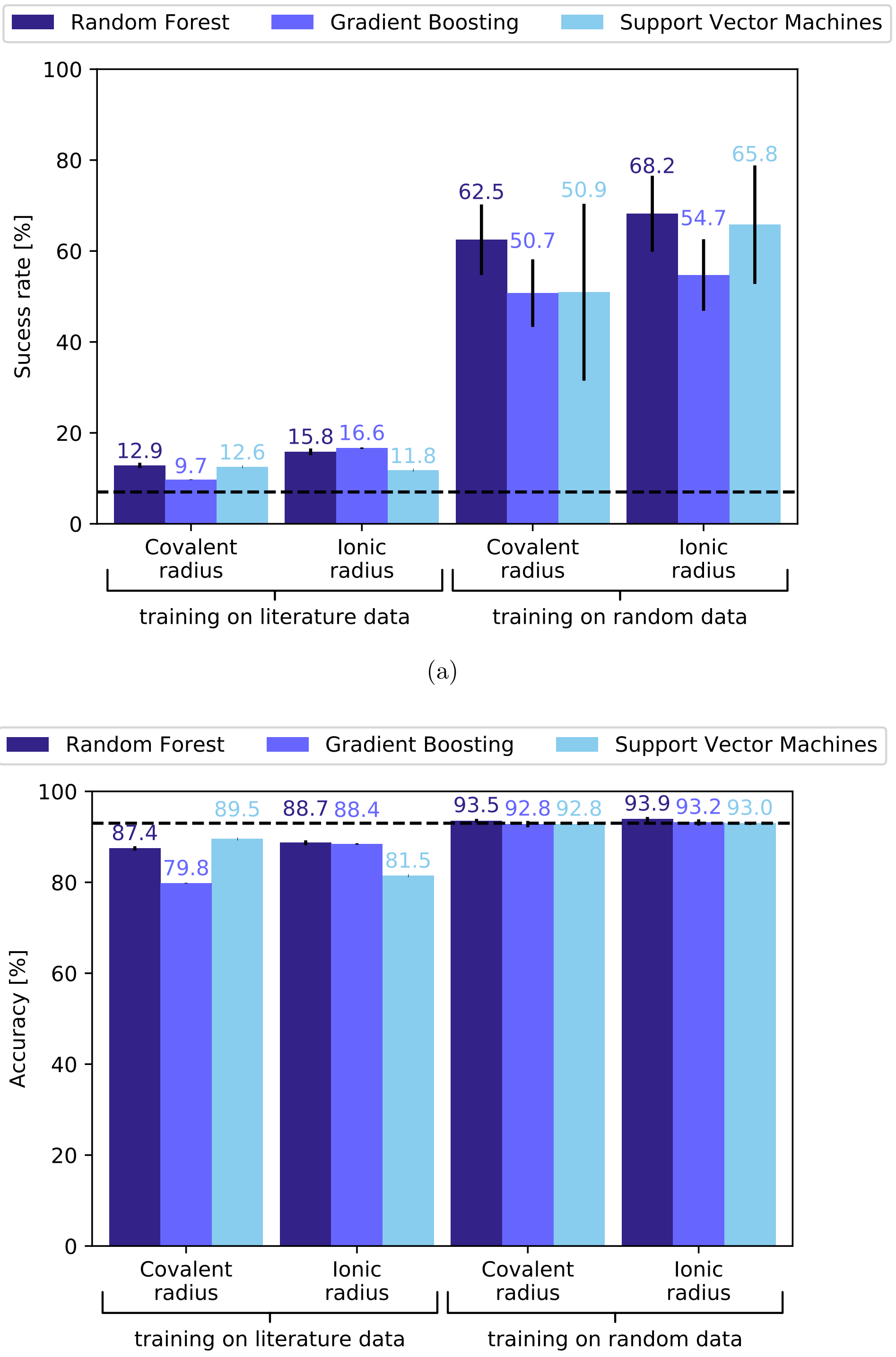
Figure 6.3. Schematic of the iterative search approach. X represents the "next generation size", i.e. the number of compounds that are calculated and added to the training set at each iteration (in this work, the generation size is 10 by default).

with higher prediction confidence are calculated first). The process is then repeated until the desired number of new discoveries is reached or the computational time is exhausted. A schematic of the iterative search approach is given in Figure 6.3. Even if adaptive design strategies[80,87,197,198] and active search[196] approaches lead to faster searches in the long term (i.e. after several iterations), our greedy iterative search approach shows better results for short term discovery.

6.3. Results and Discussion

6.3.1. Single Iteration Approach

The results obtained while using a single iteration approach are summarized in Figure 6.4. We remind the reader that the single iteration approach consists of training a single model on a fixed and known training set (where both the features and target properties are known) then use the same model to predict all remaining unknown compounds (where the target properties are not known) at once.[81,195]


Random Forest
Gradient Boosting
Support Vector Machines
Sucess rate [%]
100
80
60
40
20
0
12.9
9.7
12.6
15.8
16.6
11.8
62.5
50.7
50.9
68.2
54.7
65.8
Covalent radius
Ionic radius
Covalent radius
Ionic radius
training on literature data
training on random data
Random Forest
Gradient Boosting
Support Vector Machines
Accuracy [%]
100
80
60
40
20
0
87.4
79.8
89.5
88.7
88.4
81.5
93.5
92.8
92.8
93.9
93.2
93.0
Covalent radius
Ionic radius
Covalent radius
Ionic radius
training on literature data
training on random data

(a)

(b)

Figure 6.4. Success rate and accuracy ((a) and (b), respectively) of random forest, gradient boosting and support vector machines algorithms for different attributes and training sets. The first 6 columns represent training sets taken from the literature (Figure 6.2 (a)). The last 6 columns represent algorithms trained on a random sampling of the phase space (Figure 6.2 (b)). The numbers represent the height of each bar. The black vertical lines are the standard deviations for each column. The dashed black line is the random guessing success rate (7% success rate and 93% accuracy).

Training on Literature Data. The first 6 bars of Figure 6.4 (a) show the success rate of algorithms trained on the literature dataset (173 perovskites and 170 non-perovskites, see Figure 6.2 (a)). We can see that we predict stable compounds approximately twice as well as random guessing. As the training set is biased towards stable compounds (173/343 = 50% as opposed to 7% from the overall dataset), the algorithms predict a large amount of false positive (FP = 378.26, see Table 6.2) which impacts negatively the success rate and accuracy of such approach. In addition, the effect of covalent versus ionic radius is visible. As perovskites structures are mostly ionic compounds, choosing ionic radii yield to, as expected, better results (20% increase for random forest).

In terms of accuracy (Figure 6.4 (b)), training on literature data performs close to, but not better, than guessing non-perovskite for every compound. Even if it might look concerning at first glance, we argue that this problem is inherent with dataset containing a small fraction of stable compounds, which is typically what is found in nature.[62] Furthermore, achieving 93% accuracy by predicting non-perovskite for every compound does not help predicting new perovskites.

Training on Randomly Selected Compounds. The 6 rightmost columns of Figure 6.4 (a) show the success rate of algorithm trained on randomly selected data. In this

Table 6.2. Comparison of machine learning and DFT predictions for the random forest algorithm using ionic radii as attributes for 2 types of training set: blue is training on the literature training set and red is training on randomly selected data. Numbers are averaged over 100 runs.

		Machine learning predicts			
		perovskite	non-perovskite	perovskite	non-perovskite
DFT predicts	perovskite	70.91	61.09	82.84	197.83
	non-perovskite	378.26	3371.74	43.71	3557.62

case, we randomly select 343 compounds (the same number of compounds found in the literature) out of the 4,225 compositions, calculate them with DFT and train a model on those randomly selected compounds. By doing so, we were able to drastically improve the success rate of the algorithm to about 9-10 times the random guessing rate. It also appears that training on randomly selected data is the only way to reach the accuracy of random guessing. This shows the importance of having a training set that is as close as possible to the overall training set that we are trying to predict as it reduces the number of false positive predictions (FP = 43.71 see Table 6.2). Furthermore, training on randomly selected data removes the need for literature data allowing to explore crystal structure that are mostly unknown experimentally. Not requiring any literature data also saves some time required to assemble a training dataset from the literature and is not prone to error due to spurious experimental data.

Different Types of Algorithm. Figure 6.4 shows that, in the case of our perovskite dataset and our chosen features, random forest performs the best out of the chosen algorithms, both in terms of success rate and accuracy. It is hard to generalize this statement for other dataset types as algorithms perform differently on different kinds of data.[204–206] For instance, SVM models may work better when data are more easily linearly-separable.

As random forest performs better for our application than gradient boosting and they are both decision trees ensemble methods, we chose to keep using only random forest for all further plots.

Support vector machines offers the advantage of being deterministic meaning that we only need to train a model once, potentially saving some time. However, compared to DFT calculations time, ML timing is often negligible. Despite being SVM, there is an error bar on the last column of Figure 6.4. This is a result of training the algorithm on 100 different training set selected randomly. Having different training sets also explains why the error bar for the RF random training is larger than RF models trained on literature data where the training sets are always the same. We note that is some cases, SVM algorithms predict no stable compounds (i.e. TP = FP = 0), in this case, we choose to ignore this run all together.

Discovery Rate. As alluded to in the methodology section, the relevant quantity for materials discovery is the discovery rate i.e. the number of compounds predicted to be perovskites compared to the total number of DFT calculations performed. Therefore, in subsequent plots, we chose to plot number of compounds found as a function of the number of DFT calculations done for different algorithms (RF or SVM), training set type (literature or randomly selected), training set size (100, 200 or 343), machine learning technique (single iteration or iterative) and number of compounds iteratively added to the training set (generation size) for the iterative technique (X = 1, 10 or 100). All the plots that will follow have consistent line styles for the different parameters tested. Those styles are summarized in Table 6.3. In each plot, the slope of each line corresponds to the discovery rate for the specified parameters.

Table 6.3. Line features for the different parameters that are being tested.

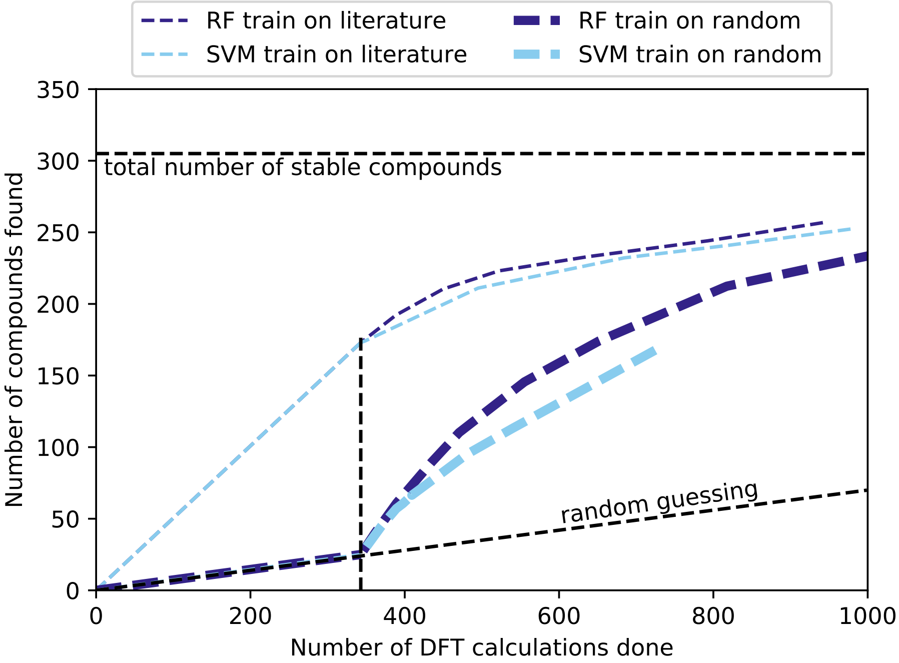
Algorithm		Training set type		Training set size		ML technique		Generation size	
RF	dark	literature	thin	100	red	single	- - -	1	-.-.-
SVM	light	random	thick	200	green	iterative	____	10	____
				343	blue			100	.....

Figure 6.5 (a) shows the number of compounds found in function of the number of calculations performed depending on the algorithm used (RF or SVM) and the type of training set (literature and random). Despite the modest success rate of the models trained on literature data ($\approx$ 15%), the high percentage of stable perovskite in the training set ($173/343 = 50\%$) makes it that literature models predict more stable perovskites compared to models trained on random selection of training data. However, most of those stable perovskites are not true discoveries as most of them were in the training set, i.e. already in the literature. In contrast, Figure 6.5 (b) reports only new compositions, i.e. compounds that were not in the training set. For the models trained on literature data, all the predictions are, by definition, new discoveries. When training on a random selection of data, some compounds in the testing set might be in the literature and thus should not be counted as new discoveries. To find out the number of newly discovered compounds, we have to calculate the probability for a compound of being not in the literature, knowing that it is stable:

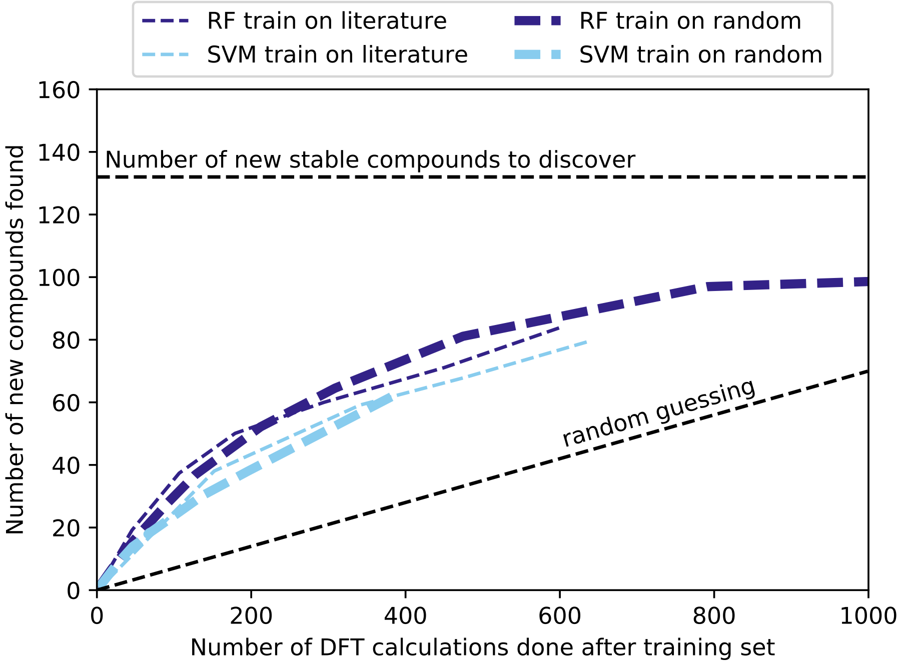
$$\text{P(not literature}|\text{stable)} = \frac{\text{P(not literature} \cap \text{stable)}}{\text{P(stable)}} = \frac{132/4225}{305/4225} = 43.3\% \tag{6.4}$$

Figure 6.5 (b) shows that the discovery rate of new compounds is similar if we train models on literature data or random selection, giving evidence that a literature training set is not always necessary.

To maximize the initial discovery rate, the order in which the DFT calculations are performed is based on the probability of being a perovskite: compounds are ranked decreasingly in function of their likeliness to form perovskites and we start calculating the compounds from the top of the list. We observe that the discovery rate is higher for the first compounds that are predicted but then slows down. This indicates that machine learning is rarely wrong when it predicts a compound to be perovskite with a high probability. After the initial high rate of discovery, the slope tapers off to reach the random guessing rate (7%). It is at this point that machine learning, without active learning, is not useful anymore. As a result, we stop the lines at that point in all plots.


RF train on literature
SVM train on literature
RF train on random
SVM train on random
Number of compounds found
0
50
100
150
200
250
300
350
total number of stable compounds
random guessing
0
200
400
600
800
1000
Number of DFT calculations done

(a)


RF train on literature
SVM train on literature
RF train on random
SVM train on random
Number of new compounds found
0
20
40
60
80
100
120
140
160
Number of new stable compounds to discover
random guessing
0
200
400
600
800
1000
Number of DFT calculations done after training set

(b)

Figure 6.5. Number of compounds calculated to be perovskites in function of the number of DFT calculations performed. (a) Considering all stable compounds and the training set. The vertical black dashed line represents the training set size (343) (b) Considering only the compounds not present in the literature. Lines are stopped when the discovery rate goes below the random guessing rate (7%).

6.3.2. Iterative Search Approach

Now that we looked at the best strategy for selecting the training set as well as the algorithm and since the discovery rate is higher right after training a model, we explore how to best use these parameters in an iterative search approach where compounds are being calculated and included in the training set continuously. The hope here is to retain the high discovery rate after each training iterations. In this section, we present results of machine learning algorithm using an iterative search approach as explained in Figure 6.3.

Single Iteration Versus Iterative Method. Figure 6.6 shows the comparison of discovery rate between the single iteration and the iterative search approach, for both random forest and support vector machines as well as both types of training set (literature and random). Similarly to the single iteration approach, RF performs better than SVM across the board. We note that as we train and predict models several times with the iterative search approach, no time is saved by using SVM over RF. In addition, the discovery rate trends between training models on literature and random data are the same as the single iteration approach.

The most important observation is that the iterative search approach is always better than the single iteration approach, regardless of the algorithm or training set type. This puts the emphasis on the advantage of continuously training new model and can be use in autonomous frameworks where machine learning dictates which DFT calculations to run next. Our iterative search method is comparable to greedy algorithms as it focuses on finding solutions as fast as possible. In contrast to active search techniques that use look-ahead or optimal search strategies, our method does not select candidate materials with an objective of also improving the ML model. With such short-term perspective, greedy algorithms are rarely optimal to find all stable compounds. Here, we argue that such a myopic strategy is useful because machine learning is not suitable to find all compounds due to 1) discovery rate drops off rapidly and 2) the only way to guarantee that we find all compounds is to calculate them all. As such, the greediness of our approach has little drawbacks.

Training Size Effect. Figure 6.7 shows the effect of decreasing the training set size on the discovery rate for (a) RF and (b) SVM. Surprisingly, the initial discovery rate, i.e. the predictions right after the initial training set, is comparable for all training set size. Furthermore, all curves are converging to a similar value after 1000 calculations. These observations imply that machine learning can greatly help materials discovery, even with little or no literature data. The fact that all methods converge to a same discovery rate is evidence of the greediness of the iterative search method where all "easy-to-find" compounds are predicted first, leaving the remaining compounds to be found at the random guessing rate. Standard deviations over the 100 runs appear to increase with

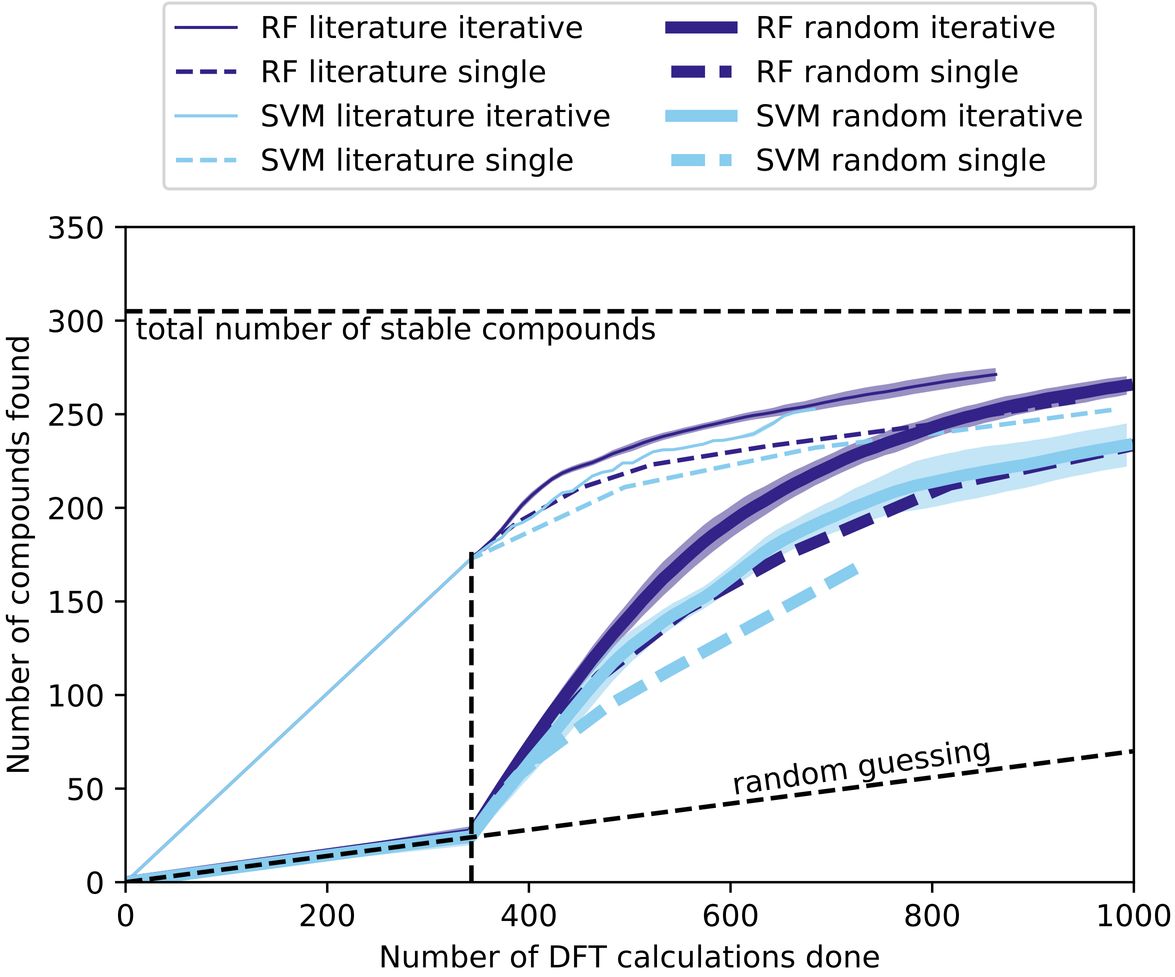
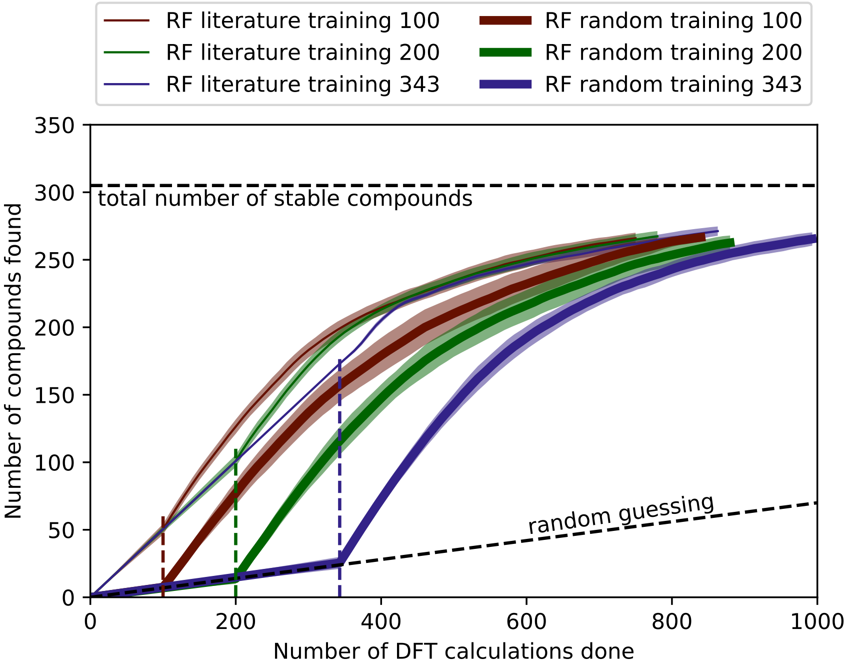
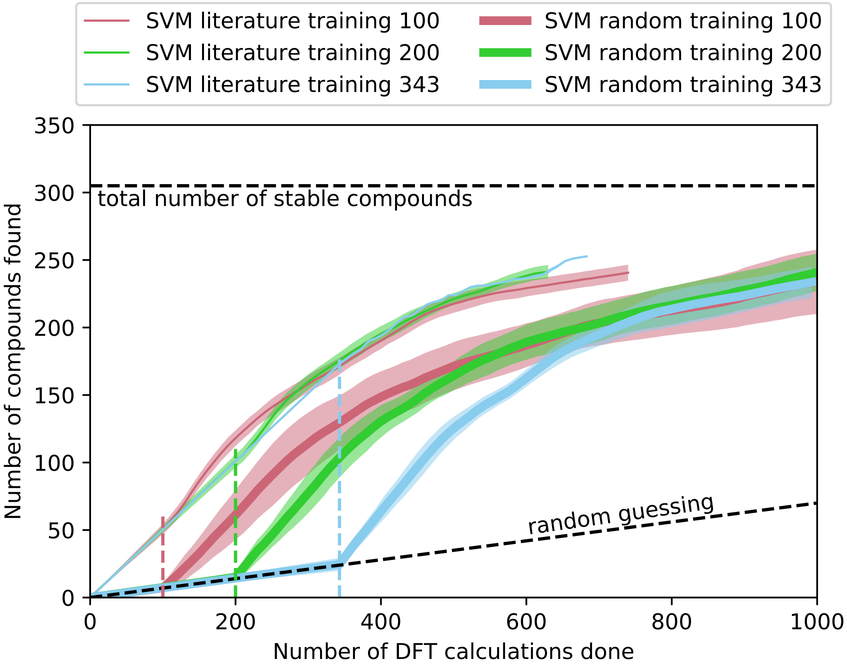


Figure 6.6. Number of compounds calculated to be perovskites in function of the number of DFT calculations performed using different machine learning technique (single iteration and iterative), different algorithm (RF and SVM) and training set type (literature and random). The vertical black dashed line represents the training set size (343). Lines are stopped when the discovery rate goes below the random guessing rate (7%).

the decrease in training set size. These fluctuations can be explained by the larger number of candidates left in the testing set when reducing the size of the training set.


RF literature training 100
RF literature training 200
RF literature training 343
RF random training 100
RF random training 200
RF random training 343
350
300
250
200
150
100
50
0
total number of stable compounds
random guessing
Number of compounds found
0
200
400
600
800
1000
Number of DFT calculations done

(a)


SVM literature training 100
SVM literature training 200
SVM literature training 343
SVM random training 100
SVM random training 200
SVM random training 343
350
300
250
200
150
100
50
0
total number of stable compounds
random guessing
Number of compounds found
0
200
400
600
800
1000
Number of DFT calculations done

(b)

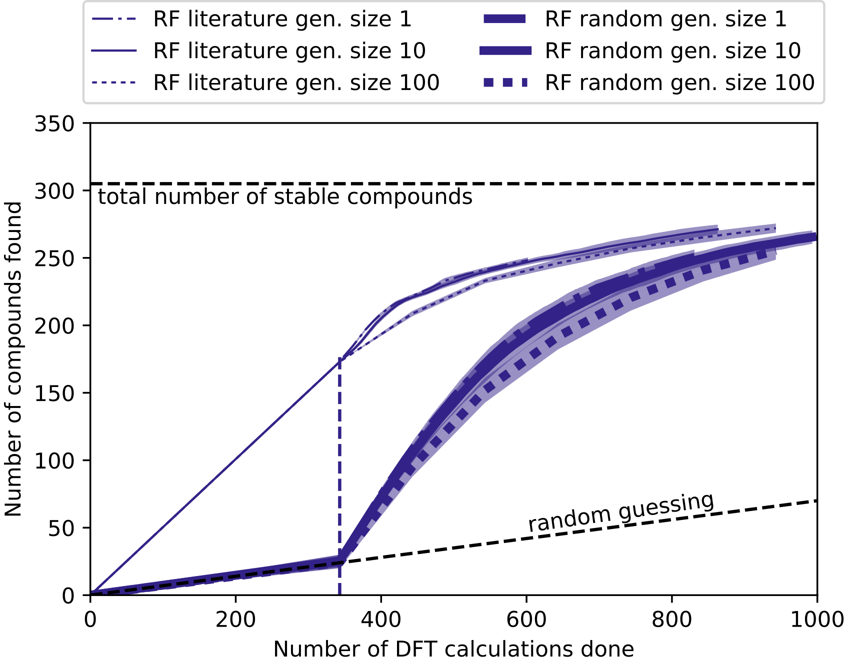
Figure 6.7. Number of compounds calculated to be perovskites in function of the number of DFT calculations performed using different training set size. The vertical colored dashed lines represent the training set size (100, 200 and 343) (a) using random forest and (b) using support vector machines. Lines are stopped when the discovery rate goes below the random guessing rate (7%).

Generation Size Effect. In Figure 6.6 and Figure 6.7, we used the iterative search approach by selecting the top 10 materials after each training iteration (i.e. generation size = 10). However, this number can be tuned. Figure 6.8 explores the effect of this parameter on the discovery rate of perovskite compounds for (a) RF and (b) SVM. It appears that a decrease in generation size increases the discovery rate. This apparent gain must be balanced by the fact that smaller generation size means less DFT calculations running in parallel and thus increase in total real time. Indeed, X = 1 essentially means that all compounds will be calculated in series which is impractical. In addition, up until now machine learning time was considered negligible compared to DFT calculation time, however, training and predict model after each compound makes the ML timing significant.

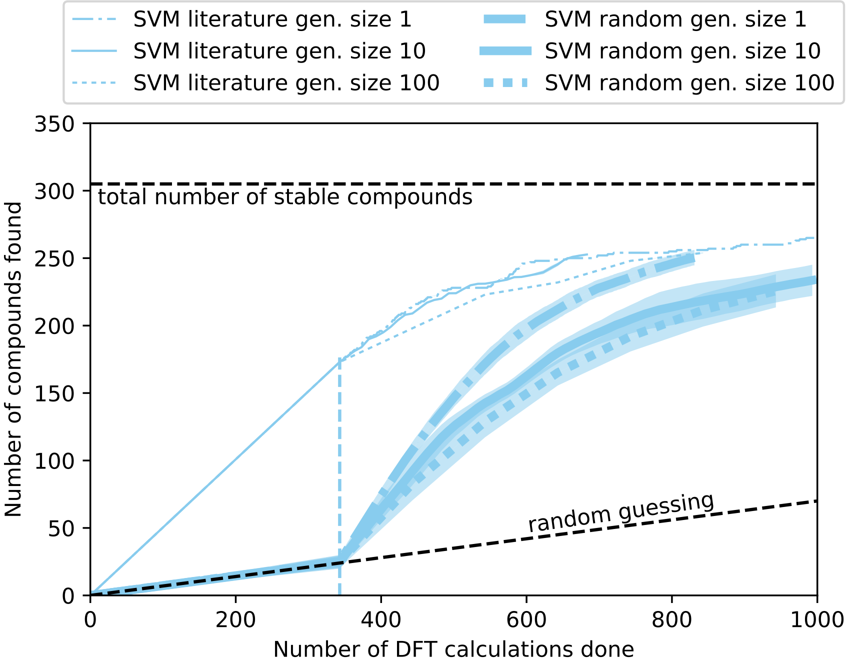
6.3.3. What Works the Best?

In this section, we discuss the optimal parameters to find new stable compounds with a certain structure. We are trying to answer three questions: 1) how to choose a training set? 2) how to build a model? And 3) how to iterate on the training set?

How to Choose a Training Set? In order to have a high success rate, the training set has to have a ratio of stable-to-non-stable compounds that is as close as possible to



(a)



(b)

Figure 6.8. Number of compounds calculated to be perovskites in function of the number of DFT calculations performed using different generation sizes. The vertical colored dashed lines represent the training set size (343) (a) using random forest and (b) using support vector machines. Lines are stopped when the discovery rate goes below the random guessing rate (7%).

the overall dataset. One way to achieve this is to randomly select compounds, calculate their stability and build a model based of those compounds. As long as there is a handful of stable compounds in the training set, larger training set does not imply faster predictions. In fact, we showed that randomly selecting 100 ABO_3 compositions, resulting in 7 perovskites and 93 non-perovskites, is enough to train a machine learning model (thick red curve in Figure 6.7).

How to Build a Model? We showed that having features that are closely related to the structure of interest, such as using ionic instead of covalent atomic radii, yields to better results. As for algorithms, ensemble methods based on decision trees, especially random forest works the best for our type of datasets, i.e. datasets that are not easily linearly-separable. We still recommend testing several different algorithms for new problems, as there is no guarantee that RF will work optimally for all projects. Finally, in the case of sparse stable compounds, the accuracy of a model is not the relevant performance metrics. Success rate or positive predictive value gives a better idea of how much DFT computer time is saved.

How to Iterate on the Training Set? Using an iterative search approach, where compounds are calculated and included in the training set outperforms the single iteration approach consisting of training a model and predicting once. For short-terms benefits, i.e. fast discovery of 80%-85% of the stable compounds, calculating the compounds with the highest prediction confidence increases drastically the initial discovery rate, right after finishing training.

What About Training on Literature Labels? Instead of training models based on DFT stability, it is tempting to train models based on literature labels, i.e. whether a

Table 6.4. Comparison between literature and DFT labels for the dataset of 343 ABO_3 compounds.

		literature labels		
		perovskite	non-perovskites	
DFT	perovskite	171	2	173
labels	non-perovskites	76	94	170
		247	96	

compound is reported as a perovskite or a different ABO_3 structure. This would save some computer time by not calculating compounds from the training set but is an unwise strategy as it creates a mismatch between the quantity we are training on (experimental stability) and the quantity we are predicting (DFT stability). As illustrated in Table 6.4, and in the case of our ABO_3 dataset, those two quantities are not equivalent. Indeed, when a compound is predicted to be a perovskite by DFT, there is good chance that it is also reported as a perovskite in the literature (171/173 = 99%). However, many perovskites reported in the literature are not predicted to be stable in the perovskite structure by DFT (76, see Table 6.4). We note here that there is a difference between DFT stability and synthesizability. However, it is not the topic of this paper and we refer the reader to other studies dealing with this matter.[117]

6.4. Summary and Conclusions

Having a large dataset of DFT calculated ABO_3 energies allowed us to test different parameters on the performance of machine learning to discover new materials. In addition to testing different algorithms such as ensemble methods and support vector machines, we investigated the effect of training set size and training set type on the discovery rate of stable perovskites. We showed that building a training set by randomly selecting compounds out of the entire search space is beneficial for the performance of the algorithm. Based on the optimal training set and algorithm, we tried an iterative search approach where unknown compounds are continuously calculated and included in the training set. We showed that this approach leads to faster materials discovery. The present machine learning comparisons can serve as a baseline for other studies aiming at discovering new materials via machine learning.

CHAPTER 7

Summary and Outlook

7.1. Summary

In this thesis, we used high-throughput density functional theory to calculate electronic and structural properties of an exhaustive list of 5,329 ABO_3 compounds. After comparing our results with the phases found in the literature, we screened for materials suitable for thermochemical water splitting based on stability and oxygen vacancy formation energy criteria. We identified 139 potential new compounds for water splitting, some of those never reported in the literature. In addition, we drew some structural maps containing both stable and unstable perovskites.

Along the way, we explained why ceria and cerium containing compounds seem to perform the best for redox reactions. For this, we looked at a source of entropy that was neglected in previous studies, the so-called onsite electronic entropy ($\Delta S_{\text{elec}}^{\text{onsite}}$). It arises from a coupling between orbital and angular spin momenta in lanthanides f-orbitals and is uniquely large for the Ce^{+4}/Ce^{+3} reduction reaction. We showed that this additional source of entropy can surpass the contribution of the vibrational entropy and configurational entropy at large oxygen off-stoichiometry. In addition, it explains the discrepancy between previous theoretical studies and experimental measurements.

With the exhaustive ABO_3 dataset at hand, we were able to test different types of machine learning techniques to understand the influences of several parameters. Our aim was to discover materials predicted to be stable with a high success rate, i.e. to have a ratio compounds stable to total number of calculations performed as high as possible. Different algorithms such as ensemble methods and support vector machines were investigated along with an iterative greedy approach aiming at maximizing the short-term performance of the algorithms. This allowed us to propose a way to do machine learning for materials discovery as efficiently as possible.

7.2. Outlook

The screening work presented in this thesis has some natural extensions that can be performed to strengthen our predictions. Aside from stability and oxygen vacancy formation energy, several other properties, such as entropy of reduction and kinetics, are crucial for thermochemical water splitting. Those quantities are more expensive to calculate and thus, have to be computed on a reduced pool of compounds. Phonons, for instance, can be calculated with density functional theory[207,208] and can be used to get a more accurate estimate of the free energy of a system. Even though the localization of oxygen vacancies has been studied experimentally and by density functional theory,[209–213] kinetics play a role in the overall performance of the system and could be further studied to sort out the new predicted materials.[43]

Experimentally, mixing perovskite seems to be a good strategy to improve the performance of water splitting materials. In particular, we can take advantage of the large entropy of reduction of the Ce^{+4}/Ce^{+3} and Tb^{+4}/Tb^{+3} redox reactions by mixing some of these elements into the perovskite crystal structure. Alternatively, searching for other crystal structures where Ce and Tb have an oxidation state of +4 might open additional novel compounds and structures to explore.

Aside from thermochemical water splitting, perovskites are used in a variety of other domains where different properties are of interest. Having a database of relaxed compounds predicted to be stable is a great starting point to start investigating other properties for different applications. For instance, detailed band structures and density of states plots are useful tools to identify materials suitable for thermoelectrics,[214] half-metals used in spintronics,[215] chemical looping[216,217] or photochemical water splitting.[6]

On the side of machine learning, the technique that we have highlighted can be used to predict the stability of many different crystal structures without spending too much time on calculating unstable compounds. Common structural prototypes that are heavily represented in experimental databases such as AB_2O_4 spinels or $ThCr_2Si_2$ can be calculated and included into the OQMD more efficiently this way. In addition, our exhaustive, publicly available dataset of ABO_3 can be used as benchmark for new machine learning techniques or algorithms deployed for materials informatics.

References

[1] International Energy Agency, *World Energy Outlook 2016*; 2016; p 684.

[2] U.S. Energy Information Administration, *Annual Energy Outlook 2017*; 2017; p 64.

[3] Abanades, S.; Charvin, P.; Flamant, G.; Neveu, P. *Energy* **2006**, *31*, 2805–2822.

[4] Wang, Z.; Roberts, R.; Naterer, G.; Gabriel, K. *International Journal of Hydrogen Energy* **2012**, *37*, 16287–16301.

[5] Khaselev, O.; Bansal, A.; Turner, J. A. *International Journal of Hydrogen Energy* **2001**, *26*, 127–132.

[6] Amouyal, E. *Solar Energy Materials and Solar Cells* **1995**, *38*, 249–276.

[7] Osterloh, F. E. *Chemistry of Materials* **2008**, *20*, 35–54.

[8] Smestad, G. P.; Steinfeld, A. *Industrial & Engineering Chemistry Research* **2012**, *51*, 11828–11840.

[9] Mclamb, N.; Sahoo, P. P.; Fuoco, L.; Maggard, P. A. *Cryst. Growth Des.* **2013**, *13*, 2322–2326.

[10] Nakamura, T. *Solar Energy* **1977**, *19*, 467–475.

[11] Roeb, M.; Neises, M.; Monnerie, N.; Call, F.; Simon, H.; Sattler, C.; Schmücker, M.; Pitz-Paal, R. *Materials* **2012**, *5*, 2015–2054.

[12] Muhich, C. L.; Evanko, B. W.; Weston, K. C.; Lichty, P.; Liang, X.; Martinek, J.; Musgrave, C. B.; Weimer, A. W. *Science (New York, N.Y.)* **2013**, *341*, 540–2.

[13] Kodama, T. *Progress in Energy and Combustion Science* **2003**, *29*, 567–597.

[14] Steinfeld, A. *Solar Energy* **2005**, *78*, 603–615.

[15] De Beni, G.; Marchetti, C. *163rd Natl Meet. Am. Chem. Soc., Div. Fuel Chem.* **1972**, 110–133.

[16] Steinfeld, A. *International Journal of Hydrogen Energy* **2002**, *27*, 611–619.

[17] Loutzenhiser, P. G.; Meier, A.; Steinfeld, A. *Materials* **2010**, *3*, 4922–4938.

[18] Abanades, S.; Charvin, P.; Lemont, F.; Flamant, G. *International Journal of Hydrogen Energy* **2008**, *33*, 6021–6030.

[19] Abanades, S. *International Journal of Hydrogen Energy* **2012**, *37*, 8223–8231.

[20] Otsuka, K.; Hatano, M.; Morikawa, A. *Inorganica Chimica Acta* **1985**, *109*, 193–197.

[21] Abanades, S.; Flamant, G. *Solar Energy* **2006**, *80*, 1611–1623.

[22] Abanades, S.; Legal, A.; Cordier, A.; Peraudeau, G.; Flamant, G.; Julbe, A. *Journal of Materials Science* **2010**, *45*, 4163–4173.

[23] Chueh, W. C.; Falter, C.; Abbott, M.; Scipio, D.; Furler, P.; Haile, S. M.; Steinfeld, A. *Science* **2010**, *330*, 1797–1801.

[24] Le Gal, A.; Abanades, S. *International Journal of Hydrogen Energy* **2011**, *36*, 4739–4748.

[25] Furler, P.; Scheffe, J. R.; Steinfeld, A. *Energy & Environmental Science* **2012**, *5*, 6098–6103.

[26] Sibieude, F.; Ducarroir, M.; Tofighi, A.; Ambriz, J. *Int. J. Hydrogen Energy* **1982**, *7*, 79–88.

[27] Allendorf, M. D.; Diver, R. B.; Siegel, N. P.; Miller, J. E. *Energy & Fuels* **2008**, *22*, 4115–4124.

[28] Kodama, T.; Nakamuro, Y.; Mizuno, T. *Journal of Solar Energy Engineering* **2006**, *128*, 3.

[29] Chueh, W. C.; Haile, S. M. *ChemSusChem* **2009**, *2*, 735–739.

[30] Chueh, W. C.; Haile, S. M. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **2010**, *368*, 3269–3294.

[31] Roeb, M.; Sattler, C. *Science (New York, N.Y.)* **2013**, *341*, 470–1.

[32] Nalbandian, L.; Evdou, A.; Zaspalis, V. *International Journal of Hydrogen Energy* **2009**, *34*, 7162–7172.

[33] McDaniel, A. H.; Miller, E. C.; Arifin, D.; Ambrosini, A.; Coker, E. N.; O'Hayre, R.; Chueh, W. C.; Tong, J. *Energy & Environmental Science* **2013**, *6*, 2424–2428.

[34] Scheffe, J. R.; Weibel, D.; Steinfeld, A. *Energy & Fuels* **2013**, *27*, 4250–4257.

[35] Yang, C.-K.; Yamazaki, Y.; Aydin, A.; Haile, S. M. *Journal of Materials Chemistry A* **2014**, *2*, 13612.

[36] Tyner, C. E.; Kolb, G. J.; Geyer, M.; Romero, M. *SolarPACES* **2001**,

[37] Haltiwanger, J. F.; Davidson, J. H.; Wilson, E. J. *Journal of Solar Energy Engineering* **2010**, *132*, 041011.

[38] Bhalla, A. S.; Guo, R.; Roy, R.; Ruyan, A. S. B.; Rustum, G. *Materials Research Innovations* **2000**, *4*, 3–26.

[39] Ishihara, T. *Perovskite Oxide for Solid Oxide Fuel Cells*; Fuel Cells and Hydrogen Energy; Springer US: Boston, MA, 2009; pp 1–296.

[40] Lufaso, M. W.; Woodward, P. M. *Acta Crystallographica Section B: Structural Science* **2001**, *57*, 725–738.

[41] Goldschmidt, V. M. *Die Naturwissenschaften* **1926**, *14*, 477–485.

[42] Stolen, S.; Bakken, E.; Mohn, C. E. *Phys. Chem. Chem. Phys.* **2006**, *8*, 429–447.

[43] Chroneos, a.; Vovk, R.; Goulatis, I.; Goulatis, L. *Journal of Alloys and Compounds* **2010**, *494*, 190–195.

[44] Kanhere, P.; Chen, Z. *Molecules* **2014**, *19*, 19995–20022.

[45] Niu, G.; Guo, X.; Wang, L. *J. Mater. Chem. A* **2015**, *3*, 8970–8980.

[46] Zhu, T.; Fowler, D. E.; Poeppelmeier, K. R.; Han, M.; Barnett, S. A. *Journal of The Electrochemical Society* **2016**, *163*, F952–F961.

[47] Kubicek, M.; Bork, A. H.; Rupp, J. L. M. *J. Mater. Chem. A* **2017**, *00*, 1–18.

[48] Schrödinger, E. *Annalen der Physik* **1926**, *384*, 361–376.
[49] Schrödinger, E. *Annalen der Physik* **1926**, *384*, 489–527.
[50] Hartree, D. R. *Mathematical Proceedings of the Cambridge Philosophical Society* **1928**, *24*, 89.
[51] Fock, V. *Zeitschrift fuer Physik* **1930**, *61*, 126–148.
[52] Hartree, D. R.; Hartree, W. *Proceedings of the Royal Society of London* **1935**, *150*, 9.
[53] Slater, J. C. *Physical Review* **1951**, *81*, 385–390.
[54] Hohenberg, P.; Kohn, W. *Physical Review* **1964**, *136*, 864–871.
[55] Kohn, W.; Sham, L. J. *Physical Review* **1965**, *140*, A1133–A1138.
[56] Perdew, J. P.; Zunger, A. *Physical Review B* **1981**, *23*, 5048–5079.
[57] Perdew, J. P.; Jackson, K. A.; Pederson, M. R.; Singh, D. J.; Fiolhais, C. *Physical Review B* **1992**, *46*, 6671–6687.
[58] Hafner, J.; Wolverton, C.; Ceder, G. *MRS Bulletin* **2006**, *31*, 659–668.
[59] Jain, A.; Hautier, G.; Moore, C. J.; Ping Ong, S.; Fischer, C. C.; Mueller, T.; Persson, K. A.; Ceder, G. *Computational Materials Science* **2011**, *50*, 2295–2310.
[60] Curtarolo, S.; Setyawan, W.; Wang, S.; Xue, J.; Yang, K.; Taylor, R. H.; Nelson, L. J.; Hart, G. L.; Sanvito, S.; Buongiorno-Nardelli, M.; Mingo, N.; Levy, O. *Computational Materials Science* **2012**, *58*, 227–235.
[61] Landis, D. D.; Hummelshoj, J. S.; Nestorov, S.; Greeley, J.; Dulak, M.; Bligaard, T.; Norskov, J. K.; Jacobsen, K. W. *Computing in Science & Engineering* **2012**, *14*, 51–57.
[62] Kirklin, S.; Saal, J. E.; Meredig, B.; Thompson, A.; Doak, J. W.; Aykol, M.; Rühl, S.; Wolverton, C. *npj Computational Materials* **2015**, *1*, 15010.
[63] Ceder, G.; Chiang, Y.-M.; Sadoway, D. R.; Aydinol, M. K.; Jang, Y.-I.; Huang, B. *Nature* **1998**, *392*, 694–696.
[64] Curtarolo, S.; Morgan, D.; Persson, K.; Rodgers, J.; Ceder, G. *Physical Review Letters* **2003**, *91*, 135503.
[65] Greeley, J.; Jaramillo, T. F.; Bonde, J.; Chorkendorff, I.; Nørskov, J. K. *Nature Materials* **2006**, *5*, 909–913.
[66] Armiento, R.; Kozinsky, B.; Fornari, M.; Ceder, G. *Physical Review B* **2011**, *84*, 014103.
[67] Kirklin, S.; Meredig, B.; Wolverton, C. *Advanced Energy Materials* **2013**, *3*, 252–262.
[68] Emery, A. A.; Saal, J. E.; Kirklin, S.; Hegde, V. I.; Wolverton, C. *Chemistry of Materials* **2016**, *28*, 5621–5634.
[69] Bergerhoff, G.; Hundt, R.; Sievers, R.; Brown, I. D. *Journal of Chemical Information and Modeling* **1983**, *23*, 66–69.
[70] Belsky, A.; Hellenbrandt, M.; Karen, V. L.; Luksch, P. *Acta Crystallographica Section B Structural Science* **2002**, *58*, 364–369.
[71] Huang, T. *19th CERN School of Computing* **1996**, 21–25.

[72] Bolton, R. J.; Hand, D. J. *Statistical Science* **2002**, *17*, 235–249.

[73] Hinton, G.; Deng, L.; Yu, D.; Dahl, G. E.; Mohamed, A.-r.; Jaitly, N.; Senior, A.; Vanhoucke, V.; Nguyen, P.; Sainath, T. N.; Kingsbury, B. *IEEE Signal Processing Magazine* **2012**, 82–97.

[74] Ramsden, J. In *Bioinformatics*; Oprea, T. I., Ed.; Computational Biology; Springer London: London, 2009; Vol. 10.

[75] Ward, L.; Wolverton, C. *Current Opinion in Solid State and Materials Science* **2016**,

[76] Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*; Springer Series in Statistics; Springer New York: New York, NY, 2009; Chapter Unsupervis, pp 1–101.

[77] Hall, M.; Frank, E.; Holmes, G.; Pfahringer, B.; Reutemann, P.; Witten, I. H. *ACM SIGKDD Explorations Newsletter* **2009**, *11*, 10.

[78] Witten, I. H.; Frank, E.; Hall, M. A. *Data Mining Practical Machine Learning Tools and Techniques*, third edit ed.; Burlington, MA : Morgan Kaufmann, 2011.

[79] Pedregosa, F. et al. *Journal of Machine Learning Research* **2012**, *12*, 2825–2830.

[80] Seko, A.; Maekawa, T.; Tsuda, K.; Tanaka, I. *Physical Review B* **2014**, *89*, 054303.

[81] Pilania, G.; Balachandran, P. V.; Gubernatis, J. E.; Lookman, T. *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials* **2015**, *71*, 507–513.

[82] Ward, L.; Agrawal, A.; Choudhary, A.; Wolverton, C. *npj Computational Materials* **2016**, *2*, 16028.

[83] Faber, F.; Lindmaa, A.; Von Lilienfeld, O. A.; Armiento, R. *International Journal of Quantum Chemistry* **2015**, *115*, 1094–1101.

[84] Ghiringhelli, L. M.; Vybiral, J.; Levchenko, S. V.; Draxl, C.; Scheffler, M. *Physical Review Letters* **2015**, *114*, 1–5.

[85] Rajan, K. *Annu. Rev. Mater. Res* **2015**, *45*, 153–69.

[86] Agrawal, A.; Choudhary, A. *APL Materials* **2016**, *4*, 053208.

[87] Lookman, T.; Alexander, F. J.; Bishop, A. R. *APL Materials* **2016**, *4*.

[88] Kresse, G.; Furthmüller, J. *Physical Review B* **1996**, *54*, 11169–11186.

[89] Kresse, G.; Furthmüller, J. *Computational Materials Science* **1996**, *6*, 15–50.

[90] Kresse, G.; Joubert, D. *Physical Review B* **1999**, *59*, 1758–1775.

[91] Perdew, J. P.; Burke, K.; Ernzerhof, M. *Physical Review Letters* **1996**, *77*, 3865–3868.

[92] Dudarev, S. L.; Botton, G. A.; Savrasov, S. Y.; Humphreys, C. J.; Sutton, A. P. *Physical Review B* **1998**, *57*, 1505–1509.

[93] Wang, L.; Maxisch, T.; Ceder, G. *Physical Review B* **2006**, *73*, 195107.

[94] Lee, Y.-L.; Kleis, J.; Rossmeisl, J.; Morgan, D. *Physical Review B* **2009**, *80*, 224101.

[95] Stevanović, V.; Lany, S.; Zhang, X.; Zunger, A. *Physical Review B* **2012**, *85*, 115104.

[96] Saal, J. E.; Kirklin, S.; Aykol, M.; Meredig, B.; Wolverton, C. *JOM* **2013**, *65*, 1501–1509.

[97] Akbarzadeh, A. R.; Ozolins, V.; Wolverton, C. *Advanced Materials* **2007**, *19*, 3233–3239.

[98] Curnan, M. T.; Kitchin, J. R. *The Journal of Physical Chemistry C* **2014**, *118*, 28776–28790.

[99] Deml, A. M.; Stevanović, V.; Holder, A. M.; Sanders, M.; O'Hayre, R.; Musgrave, C. B. *Chemistry of Materials* **2014**, *26*, 6595–6602.

[100] Kuo, J.; Anderson, H.; Sparlin, D. *Journal of Solid State Chemistry* **1989**, *83*, 52–60.

[101] Nowotny, J.; Rekas, M. *Journal of the American Ceramic Society* **2005**, *81*, 67–80.

[102] Mizusaki, J.; Yoshihiro, M.; Yamauchi, S.; Fueki, K. *Journal of Solid State Chemistry* **1985**, *58*, 257–266.

[103] Mizusaki, J.; Mima, Y.; Yamauchi, S.; Fueki, K.; Tagawa, H. *Journal of Solid State Chemistry* **1989**, *80*, 102–111.

[104] O'Keefe, M.; Brese, N. E. *Journal of the American Chemical Society* **1991**, *113*, 3226–3229.

[105] Ong, S. P.; Richards, W. D.; Jain, A.; Hautier, G.; Kocher, M.; Cholia, S.; Gunter, D.; Chevrier, V. L.; Persson, K. A.; Ceder, G. *Computational Materials Science* **2013**, *68*, 314–319.

[106] Shannon, R. D.; Prewitt, C. T. *Acta Crystallographica Section B Structural Crystallography and Crystal Chemistry* **1969**, *25*, 925–946.

[107] Shannon, R. D. *Acta Crystallographica Section A* **1976**, *32*, 751–767.

[108] Seshadri, R.; Basu, R. Periodic table of the elements. 2002.

[109] Emery, A. A. 5329 Perovskites. 2017; `http://dx.doi.org/10.6084/m9.figshare.4833587`.

[110] Roth, R. *Journal of Research of the National Bureau of Standards* **1957**, *58*, 75–88.

[111] Curtarolo, S.; Morgan, D.; Ceder, G. *Calphad: Computer Coupling of Phase Diagrams and Thermochemistry* **2005**, *29*, 163–211.

[112] Hautier, G.; Ong, S. P.; Jain, A.; Moore, C. J.; Ceder, G. *Physical Review B* **2012**, *85*.

[113] Lejaeghere, K. et al. *Science* **2016**, *351*.

[114] Haas, P.; Tran, F.; Blaha, P. *Physical Review B - Condensed Matter and Materials Physics* **2009**, *79*, 1–10.

[115] Shishkin, M.; Kresse, G. *Physical Review B* **2007**, *75*, 235102.

[116] Perdew, J. P. *International Journal of Quantum Chemistry* **1986**, *30*, 451–451.

[117] Sun, W.; Dacek, S. T.; Ong, S. P.; Hautier, G.; Jain, A.; Richards, W.; Gamst, A. C.; Persson, K. A.; Ceder, G. *Science Advances* **2016**, *2*.

[118] Schlapbach, L.; Züttel, A. *Nature* **2001**, *414*, 353–358.

[119] Singhal, S. *Solid State Ionics* **2002**, *152-153*, 405–410.

[120] von Spakovsky, M.; Olsommer, B. *Energy Conversion and Management* **2002**, *43*, 1249–1257.

[121] Weber, A.; Ivers-Tiffée, E. *Journal of Power Sources* **2004**, *127*, 273–283.

[122] Koroneos, C.; Dompros, A.; Roumbas, G.; Moussiopoulos, N. *International Journal of Hydrogen Energy* **2004**, *29*, 1443–1450.

[123] Momirlan, M.; Veziroglu, T. *International Journal of Hydrogen Energy* **2005**, *30*, 795–802.

[124] Romero, M.; Steinfeld, A. *Energy & Environmental Science* **2012**, *5*, 9234–9245.

[125] Meredig, B.; Wolverton, C. *Physical Review B* **2009**, *80*, 245119.

[126] Meredig, B.; Wolverton, C. *Physical Review B* **2011**, *83*, 239901.

[127] Demont, A.; Abanades, S.; Beche, E. *The Journal of Physical Chemistry C* **2014**, *118*, 12682–12692.

[128] McDaniel, A.; Ambrosini, A.; Coker, E.; Miller, J.; Chueh, W.; O'Hayre, R.; Tong, J. *Energy Procedia* **2014**, *49*, 2009–2018.

[129] Curtarolo, S.; Setyawan, W.; Hart, G. L.; Jahnatek, M.; Chepulskii, R. V.; Taylor, R. H.; Wang, S.; Xue, J.; Yang, K.; Levy, O.; Mehl, M. J.; Stokes, H. T.; Demchenko, D. O.; Morgan, D. *Computational Materials Science* **2012**, *58*, 218–226.

[130] Castelli, I. E.; Olsen, T.; Datta, S.; Landis, D. D.; Dahl, S.; Thygesen, K. S.; Jacobsen, K. W. *Energy & Environmental Science* **2012**, *5*, 5814–5819.

[131] Castelli, I. E.; Landis, D. D.; Thygesen, K. S.; Dahl, S.; Chorkendorff, I.; Jaramillo, T. F.; Jacobsen, K. W. *Energy & Environmental Science* **2012**, *5*, 9034–9043.

[132] Barber, C. B.; Dobkin, D. P.; Huhdanpaa, H. *ACM Transactions on Mathematical Software* **1996**, *22*, 469–483.

[133] Körbel, S.; Marques, M. A. L.; Botti, S. *J. Mater. Chem. C* **2016**, *4*, 3157–3167.

[134] Zhou, J.-S.; Goodenough, J. B. *Physical Review B* **2008**, *77*, 132104.

[135] Woodward, P. M. *Acta Crystallographica Section B Structural Science* **1997**, *53*, 32–43.

[136] Zhou, J.-S.; Alonso, J. A.; Pomjakushin, V.; Goodenough, J. B.; Ren, Y.; Yan, J.-Q.; Cheng, J.-G. *Physical Review B* **2010**, *81*, 214115.

[137] El-Mellouhi, F.; Brothers, E. N.; Lucero, M. J.; Bulik, I. W.; Scuseria, G. E. *Physical Review B* **2013**, *87*, 035107.

[138] Giaquinta, D. M.; zur Loye, H.-C. *Chemistry of Materials* **1994**, *6*, 365–372.

[139] Li, C.; Soh, K. C. K.; Wu, P. *Journal of Alloys and Compounds* **2004**, *372*, 40–48.

[140] Zhang, H.; Li, N.; Li, K.; Xue, D. *Acta Crystallographica Section B Structural Science* **2007**, *63*, 812–818.

[141] Huan, T. D.; Amsler, M.; Marques, M. A. L.; Botti, S.; Willand, A.; Goedecker, S. *Physical Review Letters* **2013**, *110*, 135502.

[142] Muller, O.; Roy, R. *The Major Ternary Structural Families*; Springer- Verlag: New York, Heidelberg, Berlin, 1974; pp 1–487.

[143] Kumar, A.; Verma, A. S.; Bhardwaj, S. R. *The Open Applied Physics Journal* **2008**, *1*, 11–19.

[144] Castelli, I. E.; Jacobsen, K. W. *Modelling and Simulation in Materials Science and Engineering* **2014**, *22*, 055007.

[145] Chiang, Y.-M.; Lavik, E.; Blom, D. *Nanostructured Materials* **1997**, *9*, 633–642.

[146] Yang, Z.; Luo, G.; Lu, Z.; Hermansson, K. *The Journal of Chemical Physics* **2007**, *127*, 074704.

[147] Murgida, G. E.; Ferrari, V.; Ganduglia-Pirovano, M. V.; Llois, A. M. *Physical Review B* **2014**, *90*, 115120.

[148] Feng, Z. a.; El Gabaly, F.; Ye, X.; Shen, Z.-X.; Chueh, W. C. *Nature Communications* **2014**, *5*, 4374.

[149] Hansen, H. A.; Wolverton, C. *The Journal of Physical Chemistry C* **2014**, *118*, 27402–27414.

[150] Otsuka, K.; Hatano, M.; Morikawa, A. *Journal of Catalysis* **1983**, *79*, 493–496.

[151] Tuller, H.; Nowick, A. *Journal of The Electrochemical Society* **1979**, *126*, 209.

[152] Yao, H.; Yu Yao, Y. *Journal of Catalysis* **1984**, *86*, 254–265.

[153] Kašpar, J.; Fornasiero, P.; Graziani, M. *Catalysis Today* **1999**, *50*, 285–298.

[154] Mogensen, M.; Sammes, N.; Tompsett, G. *Solid State Ionics* **2000**, *129*, 63–94.

[155] Gandhi, H.; Graham, G.; McCabe, R. *Journal of Catalysis* **2003**, *216*, 433–442.

[156] Sharma, S.; Hilaire, S.; Vohs, J.; Gorte, R.; Jen, H.-W. *Journal of Catalysis* **2000**, *190*, 199–204.

[157] Steele, B. *Solid State Ionics* **2000**, *129*, 95–110.

[158] Steele, B. C. H.; Heinzel, A. *Nature* **2001**, *414*, 345–352.

[159] Navrotsky, A. *Journal of Materials Chemistry* **2010**, *20*, 10577.

[160] Fu, Q.; Weber, A.; Flytzani-Stephanopoulos, M. *Catalysis Letters* **2001**, *77*, 87–95.

[161] Inaba, H.; Tagawa, H. *Solid State Ionics* **1996**, *83*, 1–16.

[162] Park, S.; Vohs, J. M.; Gorte, R. J. *Nature* **2000**, *404*, 265–267.

[163] Deluga, G. A.; Salge, J.; Schmidt, L.; Verykios, X. *Science* **2004**, *303*, 993–997.

[164] Schneider, J. J.; Naumann, M.; Schäfer, C.; Brandner, A.; Hofmann, H. J.; Claus, P. *Beilstein Journal of Nanotechnology* **2011**, *2*, 776–784.

[165] Sun, C.; Li, H.; Chen, L. *Energy & Environmental Science* **2012**, *5*, 8475.

[166] Panlener, R. J.; Blumenthal, R. N.; Garnier, J. E. *J. Phys. Chem. Solids* **1975**, *36*, 1213–1222.

[167] Nakamura, T. *Journal of solid state chemistry* **1981**, *240*, 234–240.

[168] Bevan, D.; Kordis, J. *Journal of Inorganic and Nuclear Chemistry* **1964**, *26*, 1509–1523.

[169] Grieshammer, S.; Zacherle, T.; Martin, M. *Physical chemistry chemical physics : PCCP* **2013**, *15*, 15935–42.

[170] Gopal, C. B.; van de Walle, A. *Physical Review B* **2012**, *86*, 134117.

[171] Russell, H. N.; Saunders, F. A. *The Astrophysical Journal* **1925**, *61*, 38.

[172] Van Vleck, J. H. *Physical Review* **1932**, *41*, 208–215.

[173] Zhou, F.; Åberg, D. *Physical Review B* **2016**, *93*, 085123.

[174] Kramida, A.; Ralchenko, Y.; Reader, J.; Team, N. A. NIST Atomic Spectra Database (ver. 5.3). 2015; `https://www.nist.gov/pml/atomic-spectra-database`.

[175] Yen, W. M. *Physics of the Solid State* **2005**, *47*, 1393.

[176] Myers, C. E.; Graves, D. T. *Journal of Chemical & Engineering Data* **1977**, *22*, 440–445.

[177] Bhosale, R.; Kumar, A.; AlMomani, F. *International Journal of Photoenergy* **2016**, *2016*, 1–9.

[178] Walsh, W. M.; Jeener, J.; Bloembergen, N. *Physical Review* **1965**, *139*, A1338–A1350.

[179] Zacherle, T.; Schriever, A.; De Souza, R. A.; Martin, M. *Physical Review B - Condensed Matter and Materials Physics* **2013**, *87*, 1–11.

[180] Furrer, A.; Podlesnyak, A.; Frontzek, M.; Sashin, I.; Embs, J. P.; Mitberg, E.; Pomjakushina, E. *Physical Review B* **2014**, *90*, 064426.

[181] Tuller, H.; Nowick, A. *Journal of Physics and Chemistry of Solids* **1977**, *38*, 859–867.

[182] Hachmann, J.; Olivares-Amaya, R.; Atahan-Evrenk, S.; Amador-Bedolla, C.; Sanchez-Carrera, R. S.; Gold-Parker, A.; Vogt, L.; Brockway, A. M.; Aspuru-Guzik, A. *The Journal of Physical Chemistry Letters* **2011**, *2*, 2241–2251.

[183] Hautier, G.; Fischer, C. C.; Jain, A.; Mueller, T.; Ceder, G. *Chemistry of Materials* **2010**, *22*, 3762–3767.

[184] Faber, F. A.; Lindmaa, A.; Von Lilienfeld, O. A.; Armiento, R. *Physical Review Letters* **2016**, *117*, 2–7.

[185] Kirklin, S.; Saal, J. E.; Hegde, V. I.; Wolverton, C. *Acta Materialia* **2016**, *102*, 125–135.

[186] Kalidindi, S. R.; De Graef, M. *Annual Review of Materials Research* **2015**, *45*, 171–193.

[187] Hill, J.; Mulholland, G.; Persson, K.; Seshadri, R.; Wolverton, C.; Meredig, B. *MRS Bulletin* **2016**, *41*, 399–409.

[188] Fischer, C. C.; Tibbetts, K. J.; Morgan, D.; Ceder, G. *Nature Materials* **2006**, *5*, 641–646.

[189] Hautier, G. *Topics in current chemistry*; 2013.

[190] Pilania, G.; Gubernatis, J. E.; Lookman, T. *Physical Review B - Condensed Matter and Materials Physics* **2015**, *91*, 1–13.

[191] Chatterjee, S.; Murugananth, M.; Bhadeshia, H. K. D. H. *Materials Science and Technology* **2007**, *23*, 819–827.

[192] de Jong, M.; Chen, W.; Notestine, R.; Persson, K.; Ceder, G.; Jain, A.; Asta, M.; Gamst, A. *Scientific Reports* **2016**, *6*, 34256.

[193] Meredig, B.; Agrawal, A.; Kirklin, S.; Saal, J. E.; Doak, J. W.; Thompson, A.; Zhang, K.; Choudhary, A.; Wolverton, C. *Physical Review B - Condensed Matter and Materials Physics* **2014**, *89*, 1–7.

[194] Oliynyk, A. O.; Antono, E.; Sparks, T. D.; Ghadbeigi, L.; Gaultois, M. W.; Meredig, B.; Mar, A. *Chemistry of Materials* **2016**, *28*, 7324–7331.

[195] Pilania, G.; Balachandran, P. V.; Kim, C.; Lookman, T. *Frontiers in Materials* **2016**, *3*, 1–7.

[196] Garnett, R.; Gärtner, T.; Vogt, M.; Bajorath, J. *Journal of Computer-Aided Molecular Design* **2015**, *29*, 305–314.

[197] Balachandran, P. V.; Xue, D.; Theiler, J.; Hogden, J.; Lookman, T. *Scientific Reports* **2016**, *6*, 19660.

[198] Xue, D.; Balachandran, P. V.; Hogden, J.; Theiler, J.; Xue, D.; Lookman, T. *Nature Communications* **2016**, *7*, 11241.

[199] Pilania, G.; Wang, C.; Jiang, X.; Rajasekaran, S.; Ramprasad, R. *Scientific reports* **2013**, *3*, 2810.

[200] Wolfram's element properties. `http://reference.wolfram.com/mathematica/note/ElementDataSourceInformation.html`.

[201] Breiman, L. *Machine Learning* **2001**, *45*, 5–32.

[202] Freidman, J. H. *Institue of Mathematical Statistics* **2001**, *29*, 1189–1232.

[203] Chang, C.-c.; Lin, C.-j. *ACM Transactions on Intelligent Systems and Technology (TIST)* **2013**, *2*, 1–39.

[204] Liaw, A.; Wiener, M. *R news* **2002**, *2*, 18–22.

[205] Statnikov, A.; Wang, L.; Aliferis, C. *BMC Bioinformatics* **2008**, *9*, 319.

[206] Liu, M.; Wang, M.; Wang, J.; Li, D. *Sensors and Actuators, B: Chemical* **2013**, *177*, 970–980.

[207] Baroni, S.; de Gironcoli, S.; Dal Corso, A.; Giannozzi, P. *Reviews of Modern Physics* **2001**, *73*, 515–562.

[208] van de Walle, A.; Asta, M.; Ceder, G. *Calphad* **2002**, *26*, 539–553.

[209] Sayle, T.; Parker, S.; Catlow, C. *Surface Science* **1994**, *316*, 329–336.

[210] Fabris, S.; Vicario, G.; Balducci, G.; de Gironcoli, S.; Baroni, S. *The Journal of Physical Chemistry B* **2005**, *109*, 22860–22867.

[211] Nolan, M.; Fearon, J. E.; Watson, G. W. *Solid State Ionics* **2006**, *177*, 3069–3074.

[212] Migani, A.; Vayssilov, G. N.; Bromley, S. T.; Illas, F.; Neyman, K. M. *Journal of Materials Chemistry* **2010**, *20*, 10535.

[213] Chueh, W. C.; McDaniel, A. H.; Grass, M. E.; Hao, Y.; Jabeen, N.; Liu, Z.; Haile, S. M.; McCarty, K. F.; Bluhm, H.; El Gabaly, F. *Chemistry of Materials* **2012**, *24*, 1876–1882.

[214] Kuroki, K.; Arita, R. *Journal of the Physical Society of Japan* **2007**, *76*, 083707.

[215] Gomonay, E. V.; Loktev, V. M. *Low Temperature Physics* **2014**, *40*, 17–35.

[216] Ryden, M.; Lyngfelt, A.; Mattisson, T.; Chen, D.; Holmen, A.; Bjorgum, E. *International Journal of Greenhouse Gas Control* **2008**, *2*, 21–36.

[217] Jing, D.; Mattisson, T.; Leion, H.; Ryden, M.; Lyngfelt, A. *International Journal of Chemical Engineering* **2013**, *2013*, 1–16.

Publications

(1) **Emery AA**, Saal JE, Kirklin S, Hegde VI, Wolverton C. High-Throughput Computational Screening of Perovskites for Thermochemical Water Splitting Applications. Chem. Mater. 28, 5621-5634 (2016).

(2) **Emery AA**, Wolverton C. High-Throughput DFT Calculations of Formation Energy, Stability and Oxygen Vacancy Formation Energy of ABO_3 Perovskites. Scientific Data (accepted)

(3) **Emery AA**, Ward LT, Wolverton C. Designing Optimal Machine Learning Method for Materials Discovery (in preparation)

(4) Naghavi SS, **Emery AA**, Hansen HA, Zhou F, Ozolins V, Wolverton C. Giant Onsite Electronic Entropy Enhances the Performance of Ceria for Water Splitting. Nature Communications (in press.)

(5) Balachandran PV, **Emery AA**, Gubernatis JE, Lookman T, Wolverton C, Zunger A. Predictions of New ABO_3 Compounds in Perovskite and Cubic Perovskite Structures: A Combined Machine Learning and High-Throughput Density Functional Theory Study (in preparation)

(6) Barcellos DR, Coury FG, **Emery AA**, Sanders M, Tong J, McDaniel A, Wolverton C, Kaufman M, O'Hayre R. Phase Identification of the Layered Perovskite $Sr_{2-x}Ce_xMnO_4$ for Solar Thermochemical Water Splitting (in preparation)